INSIGHT
Regeneration

nature

THE INTERNATIONAL WEEKLY JOURNAL OF SCIENCE

HOW TO GROW A HEALTHY LAB

WAYS TO CREATE AND MAINTAIN A FRUITFUL RESEARCH GROUP **PAGE 293**

CELL BIOLOGY

IDENTITY SWITCH

Transcription factors that change cells into neurons

PAGES 316 & 375

ATMOSPHERIC SCIENCE

CFC DECLINE SLOWS DOWN

Emissions of ozone depleter have increased since 2012

PAGES 317 & 413

COMPUTATIONAL NEUROSCIENCE

THE RIGHT DIRECTION

AI system mimics human brain's navigational skills

PAGES 313 & 429

➲ NATUREASIA.COM
17 May 2018
Vol. 557, No. 7705

THIS WEEK

EDITORIALS

INDUSTRY Aluminium breakthrough offers cleaner smelting pot **p.280**

LUNAR WARMING Walking on the Moon raised surface temperature **p.282**

FUNGUS Lethal amphibian fungus traced to Korean Peninsula **p.283**

A slow road for stem cells

The steady and careful development that has guided treatments using embryonic stem cells should be applied to therapies derived from adult stem cells, too.

Regulators in the United States last week sought an injunction to stop a Florida company selling a controversial adult-stem-cell treatment for age-related macular degeneration (AMD). The move comes after three women treated by the company in 2005 went blind. The firm, now called US Stem Cell, hailed the cells as a therapy for AMD, which causes vision to blur so much that people affected can no longer recognize faces.

In the same year, two other people with AMD received a different stem-cell treatment at a London hospital. Those patients had patches made from embryonic stem cells implanted into their retinas. The scientists behind that therapy reported on the patients' progress earlier this year: their eyesight had improved beyond expectations (L. da Cruz *et al. Nature Biotechnol.* **36**, 328–337; 2018).

The conflicting outcomes highlight a difference between many treatments that use adult stem cells and those based on embryonic stem (ES) cells. ES-cell therapies emerged from a slowly built body of knowledge on how cells should be created and implanted, whereas adult-stem-cell treatments have too often been propelled by empty promises rather than by evidence.

Research with human ES cells has been slow because it was forced to be. Since scientists first created the cells 20 years ago, they have faced restrictions on funding and a need to pass through extra review committees, because of the sensitive nature of research based on donated human embryos. The path was difficult to navigate and full of setbacks, but something good came out of it.

Those who dared to proceed were few, but they were committed. Working under intense scrutiny, they progressed steadily, even if the work was too sluggish for some. Enter adult stem cells. Scientists, clinics and companies lined up to capitalize on the opportunity; many compared the 'unethical' nature of human ES cells with the ethical choice of adult stem cells. Some advertised, without evidence, how they were harnessing the body's own power of rejuvenation. Some even kept a score card of adult-cell therapies marketed, versus zero for those from ES cells.

Adult stem cells can certainly be valuable. Bone-marrow transplants are a stem-cell therapy, and a tremendously successful one. Transplant of limbal stem cells found in the eye has fixed the corneas of hundreds of people. And last November, physicians in Italy reported using another kind of adult stem cell — epidermal — to save a German boy with a usually fatal skin disease (T. Hirsch *et al. Nature* **551**, 327–332; 2017).

But in too many other cases, progress has been crippled by a lack of any proof of efficacy — in particular, when it comes to therapies based on mesenchymal stem cells taken from a person. Too many companies seeking a quick profit have exploited lax regulatory frameworks — in the United States and elsewhere — and the needs of desperate people facing sometimes terminal illnesses. Patient need has been presented as an excuse to forgo clinical trials. Careful bench experiments that were needed to reveal how the therapies might work have not been done.

The situation could get worse. An article last week argued that crowdfunding campaigns to drum up money for treatments accelerate the dissemination of inaccurate information (J. Snyder *et al. J. Am. Med. Assoc.* **319**, 1935–1936; 2018).

Contrast that with solid scientific work that has shown, for example, how to use ES cells to derive retinal cells, pancreatic cells and dopamine-producing cells. Techniques based on years of rigorous work to characterize and develop ways of delivering these cells are now in or nearing clinical trials.

> ***"Adult-stem-cell treatments have too often been propelled by empty promises."***

Other positive examples are described in a special *Nature* Insight supplement this week (see page 321), which catalogues the growing knowledge base from experiments with ES cells that aim to treat diseases affecting the pancreas and brain. And it discusses innovative strategies, such as spurring ageing stem cells in the body to fight off disease.

Understandably, such progress can seem frustratingly slow to many patients. But speedy alternatives are more of a problem. Regulators have not been able to keep up. Last week's request for an injunction is being heralded as a turning point in the US Food and Drug Administration's (FDA's) crackdown on clinics offering unproven stem-cell therapies. But there are still hundreds more such operators in the United States alone.

Other new treatments already hitting the market, including immunotherapy and gene therapy, are also vulnerable to hype. If the FDA and other regulators are to have any chance of sifting the good from the bad, and so protecting some of the most vulnerable people, they need to pick up the pace. Meanwhile, scientists should remember the merits of doing the opposite. ■

Health check

Universities should ensure lab environments are supportive, productive and rigorous.

Directing academics has been compared to herding cats, animals that famously follow their own path and scorn instruction. So, while worrying, it's perhaps not surprising that two-thirds of lab heads who responded to a *Nature* survey this year said that they had received no training in mentoring or managing people. Yet two-thirds of these untrained senior scientists said they thought it would be useful.

They were right. Good-quality training is a key ingredient to building a successful research group. So, too, is the wider academic environment in which

LAB HEALTH
A *Nature* special issue
nature.com/collections/labhealth

researchers work. If a department or institution does not encourage collaboration, celebrate success or value solid work over flashy promotion — as well as training — then group leaders will struggle to create a healthy research culture in their own laboratories.

How institutions can help lab groups to be productive, supportive and rigorous is an essential but often-overlooked topic. To try to change that, this week we have made it the focus of a special issue. This builds on discussions between *Nature* editors, senior academics and postdocs held throughout 2017, and on a day-long conference co-sponsored by *Nature* and the University of California, Berkeley, in October 2017.

Laboratory members should feel that they are an integral part of their institutions and departments, but this is not always the case. According to the survey, one-fifth of those scientists who didn't lead research groups had a negative view of the culture or working environment in their lab. And the barrage of sexual-harassment allegations at universities around the world is sad evidence of how often institutions fail to protect junior members of communities.

Institutional support is essential during the daily grind of scientific research, not just in times of crisis. In the survey — covered in a News Feature on page 294 — research-group leaders were asked about ways departments could support them. The most common answers pointed to resources for administrative tasks, support for mentoring and managing lab members, and more use of measures of scientific productivity beyond counts of high-profile papers.

Institutions need to support lab members as well as leaders. A Comment piece on page 299 argues that institutions should implement a 'culture of structure' to give graduate students clear expectations of their progress, and to ensure contact with multiple faculty members.

But each department and institution has different needs, so how can leaders work out what support to offer and make sure it is welcome? As another Comment piece on page 297 advocates, they can collect benchmarking data on lab culture (such as student–supervisor relationships) to identify areas for improvement. They can also hold cross-lab meetings and ask faculty for explicit feedback in annual reports.

Some institutions have already taken the step of hiring staff to support community-building and rigorous research, such as scientists who help lab groups to implement quality-control practices. On page 302, we highlight how one institute created a dedicated science sustainability officer.

Examples of such innovations to boost lab health can be hard to identify. *Nature* hopes to detail more as part of continued coverage — please send examples to nature@nature.com.

Others in the research enterprise also have a part to play. Funders must stress the obligation on those who receive money to support and protect trainees. Journals can set clear requirements on how work should be reported, for example, to make sure that all authors are properly credited.

But institutions are in the strongest position to improve lab health — and that can be a delicate process when introducing important reforms to independent-minded researchers. Mandated steps will not produce real improvements if academics are not persuaded of the case for them. As anyone forced to sit through generic online training programs can attest, it's too easy to follow the letter of such laws without buying into their spirit.

> ***"Academic institutions must start striving to improve the health of their labs."***

More than edicts, making change requires many small discussions between stakeholders to air problems, and to build consensus and understanding. It also requires sustained commitment.

But academic institutions must start measuring and striving to improve the health of their labs. It is in their own long-term interest to do so. A small survey of North American postdocs found that a pleasant lab environment correlated significantly with life satisfaction, whereas their number of publications did not (go.nature.com/2rfjz6v). Departments that gain a reputation for better cultures will attract better scientists, because they frequently value the quality of colleagues and interaction more than funds and laboratory space. And, ultimately, that leads to more of the advances and discoveries on which the world turns. We at *Nature* hope that this special issue helps to stimulate a race to the top. ■

Smelting point

Industrial partnership and new technology promise a greener way to make aluminium.

The world produced more than 63,000 tonnes of aluminium last year, which went into everything from kitchen foil and cans to aircraft. The metal is lightweight, fully recyclable and surprisingly strong. And now, two leading aluminium companies say that it will soon be clean and green. But just like electric cars, making aluminium can ultimately be only as clean as its source of power.

On 10 May, the US manufacturer Alcoa joined with British–Australian firm Rio Tinto to announce a new joint venture, Elysis. Based in Montreal, Canada, it plans to roll out a low-carbon technology for smelting aluminium by 2024. The world will surely benefit if it does. Aluminium production accounts for 1% of global greenhouse-gas emissions each year — roughly equivalent to emissions from France in 2016. It's a two-step process: refine aluminium oxide powder from bauxite ore and then convert it to aluminium in smelters. The bulk of the industry's climate footprint is from the smelting process, which requires prodigious quantities of electricity. More than half of its total emissions come indirectly from the electricity production itself.

Here's the opportunity: nearly 20% of the emissions are from the production and degradation of the carbon anodes used to conduct electricity during smelting. And this is where the new process focuses.

Details are scant, but observers of the aluminium industry will not be surprised to hear that the Elysis technology focuses on a long-standing issue and involves a proprietary inert anode — probably a ceramic composite. Instead of releasing CO_2 and perfluorocarbons, it emits oxygen. This would completely eliminate the direct carbon emissions, but does require more electricity. If paired with alternative cathodes and new designs for the electrolytic cell, however, it would be possible to reduce electricity consumption.

Alcoa claims that, if fully implemented at all of the Canadian smelters, the technology would reduce emissions by an estimated 6.5 million tonnes each year. That all sounds good. But this isn't the first time that a major aluminium producer has talked up revolutionary smelting technology. Russian firm UC Rusal, for instance, has been seemingly on the cusp of developing inert anodes for several years. Alcoa's work goes back several decades and included a big push around 2000. The problem has long stymied academics and government researchers, too.

Industry deserves credit for continuing to invest in long-term research and development, and for not giving up on a difficult problem. But even if Elysis does succeed, aluminium production will still yield emissions from mining, aluminium oxide processing and transport. One thing industry can do is to ensure that companies and consumers recycle as much aluminium as possible, because it doesn't need to be smelted. But the biggest question is where producers get the electricity.

Companies have already begun locating aluminium smelters near hydroelectric facilities, which provide relatively cheap and reliable power. Alcoa has even moved some of its smelting operations from the United States to Iceland, which provides cheap geothermal electricity. In the end, Alcoa and Rio Tinto are like everybody else. They need a reliable source of low-carbon power to reduce their climate impact — and the cheaper the better. ■

WORLD VIEW *A personal take on events*


UNIVERSITY OF GLASGOW

Give every paper a read for reproducibility

I was hired to ferret out errors and establish routines that promote rigorous research, says **Catherine Winchester***.*

In 2012, I saw an advertisement for an unusual new role, and knew I was the one for the job. Leaders at the Cancer Research UK Beatson Institute, a non-profit organization in Glasgow, had created a position to weave discussion of and practices for research integrity into the daily routines of its roughly 300 researchers.

I had some ideas about rigorous research from working as a molecular biologist for almost 20 years, supervising graduate students and running a team. I think the best way to boost research quality is to discuss it often and freely. Intentional misconduct is rare, but even raising the topic of bad practices and sloppy mistakes can be sensitive — the elephant in the room that no one wants to talk about.

In addition to helping with grant submissions, my role is to support data curation and improve manuscript quality, and to provide training and a first point of contact for concerns about integrity. I thoroughly read every research manuscript that our scientists submit to journals, about 50 primary papers each year. I receive warm thanks much more often than cold shoulders.

I strive to be the friendly face behind a serious issue, raising awareness, changing behaviour and working to create an environment in which discussing research integrity is a normal, non-confrontational part of doing science. I meet everyone personally, including postdocs, PhD students and technicians — ideally, during their first week on the job — and I have integrated myself as a scientist at the institute. People expect to see me at seminars and social events.

THE BEST WAY TO BOOST RESEARCH **QUALITY** IS TO DISCUSS IT **OFTEN** AND **FREELY.**

Perhaps the most complex undertaking so far has been developing practices for curating and preserving all the data that underpin a paper, including replicates. This took about a year, working with senior faculty members, the information-technology team and another research manager. We trialled our data-archiving system with a couple of groups, implemented it across our institute for a year and then amended it on the basis of feedback. Instead of squirrelling away data in individual folders and lab books, researchers now archive all published data in a designated central drive, so that the information is accessible for the long haul. Initially, people thought the process was just extra bureaucratic work, or that it had been invented so I could police their data. Now, it has become the norm, and researchers tell me they save time and worry by having their data organized and archived.

Feedback about my reviews has been positive, especially because, as a fresh set of eyes, I can sometimes spot mistakes that someone closer to the work might not see. I've pointed out duplicated image panels, missing data and mislabelled images, among other problems. I also check manuscript texts for plagiarism, using software such as iThenticate, and check figures to look for inadvertent duplications or inappropriate manipulation. I have incorporated elements of journal checklists, including *Nature*'s technical and statistical checklist, into my reviews for reporting accuracy, experimental design and analyses. Occasionally, I suggest using a different statistical test or way of presenting data (plotting individual points rather than bar charts, for instance). The goal is to improve a manuscript's quality without trying to emulate peer review.

Checks are compulsory but informal. I promote my work as an extra pair of eyes, and I can usually complete checks in a week or two, because I have already read researchers' grant applications and have access to the data to answer most queries myself. Sometimes, I get a response that reads like a rebuttal to a peer reviewer; most often, it feels collaborative.

Training provides an informal forum in which to discuss research-integrity issues. Face-to-face, tailored training brings integrity matters into the open and provides relevant, practical guidance. For example, I run 90-minute workshops on data management, responsible image preparation, statistical considerations and avoiding plagiarism. Training is mandatory for everyone. For new principal investigators, it is one on one, whereas postdocs, graduate students, technicians and research assistants meet in groups of up to ten people. Graduate students receive academic credit; they are, after all, learning the essentials for being a scientist. I update curricula regularly to incorporate new policies and news stories, but the sessions work best when people talk about a publication issue or problem they have experienced themselves.

Five years on, we've had no retractions or serious issues with publications from our institute, and I feel that my and my colleagues' efforts have improved practice much more than would just putting a policy on a webpage. My door is always open. If someone does raise an issue, our approach allows for informal discussions and discrete enquiries before any formal investigation is initiated. This can prevent things from escalating unnecessarily, and it also takes some of the pressures off researchers who might want to raise concerns.

Could most university departments have research-integrity advisers? Yes — but these people need to have a research background in that discipline, and be embedded in researchers' day-to-day affairs. It would add to headcount, but in the long run it would save money: high-quality research is easier to build from, and misconduct investigations get pricey quickly. More importantly, they consume many investigators' lives, and damage credibility and public trust in science. Hopefully, initiatives such as this one will all go some way towards creating a culture in which more scientists are willing to talk about — and tame — the elephant in the room. ■

Catherine Winchester *is the grants- and research-integrity adviser at the Cancer Research UK Beatson Institute in Glasgow.*
e-mail: c.winchester@beatson.gla.ac.uk

SEVEN DAYS

The news in brief

MARIO TAMA/GETTY

Disaster declared for Hawaii volcano

Lava-spewing fissures created by the eruption of the Kīlauea volcano in Hawaii continue to threaten residents on the island. Since the eruption began on 3 May, volcanic activity has destroyed 37 structures, and 2,000 people have been evacuated from their homes. Scientists have warned that the volcano could soon begin ejecting rock and ash if the crater's lava lake drops to or below the level of the water table. On 11 May, US President Donald Trump approved a disaster declaration for Hawaii, freeing up federal funds to deal with the eruption.

EVENTS

Ebola outbreak

The Democratic Republic of Congo (DRC) declared an outbreak of the Ebola virus on 8 May, after two people in the northwestern Equateur province tested positive for the virus. There have been 39 suspected cases of Ebola in the province since 4 April, including 19 deaths, the World Health Organization said on 14 May. This is the DRC's ninth Ebola epidemic since 1976. The last outbreak occurred in 2017. In 2014, the virus killed more than 11,000 people across West Africa.

Nuclear withdrawal

US President Donald Trump announced on 8 May that he will pull the United States out of the Iran nuclear deal. Researchers say that the move will hamper efforts to establish scientific collaborations between the two countries. Under the 2015 Joint Comprehensive Plan of Action, Iran agreed to scale back its nuclear programme and allow inspections of its facilities, in exchange for the removal of economic sanctions imposed by the United States and its allies. Iranian scientists expanded working relationships with their European counterparts in areas such as nuclear safety and security. But efforts to set up collaborations between US and Iranian researchers — including a project to steer Iran towards a safe and secure nuclear-energy programme — will probably come to a halt. See page 287 for more.

Nuclear dismantling

North Korea will close its nuclear test site at Mount Mantap, the country announced in the state media on 12 May. International journalists will be invited to witness the dismantling ceremony, to be held in late May. Satellite-imagery analysis published on 14 May by the independent website 38 North indicates that some buildings and equipment have already been removed from the mountain location near Punggye-ri. North Korean leader Kim Jong-un has previously said the site is no longer needed now that the nation has completed development of its nuclear weapons. He and US President Donald Trump are set to meet in Singapore on 12 June to discuss denuclearization of the Korean Peninsula.

FUNDING

R&D refund capped

The latest Australian budget, presented on 8 May, has proposed a cap on the tax refund that companies can claim if they invest in research and development (R&D). The government said that the R&D Tax Incentive had cost two-thirds more than was forecast when it was implemented in 2011. The budget proposes that the scheme be capped at Aus$4 million (US$3 million) a year for companies with annual turnovers of less than Aus$20 million. Spending on clinical trials will be exempt from the cap, much to the relief of the biotechnology industry. See page 290 for more.

South African cash

South Africa's latest science budget, approved on 9 May, cut support for its main research-funding body. The National Research Foundation (NRF) was allocated 904.8 million rands (US$73 million) for 2018–19, down 2% from 926 million rands in 2017–18. The NRF's budget has declined by 10% since 2012; last year, the agency slashed funds for a prestigious grant programme, causing some of the country's leading academics to lose up to 90% of their cash. The budget of the department of science and technology, which oversees the NRF, rose from 7.5 billion to 7.79 billion rands, a below-inflation increase. But the department says it will maintain support for the Square Kilometre Array, a project to build a giant radio telescope in South Africa and Australia. It has earmarked 709 million rands for the project in the next financial year.

SPACE

Chinese satellite

China launched its latest Earth-monitoring satellite on 9 May at the Taiyuan Satellite Launch

Centre in Shanxi province. The craft, called Gaofen-5, will boost China's capabilities in atmospheric observation. Four of its six instruments will monitor air pollution and greenhouse gases. The remaining two will survey the land and the sea and include, according to media reports, China's first hyperspectral payload, which is designed to capture high-resolution spectral imagery. The satellite is the tenth to go into orbit in the Gaofen series, whose varying instruments are intended to provide high-resolution monitoring of the entire Earth by 2020.

CONSERVATION

Rodent-free island

The sub-Antarctic island of South Georgia has declared itself free of invasive rodents, after a decade-long eradication effort. The project, which cost £10 million (US$14 million), is the biggest such programme ever undertaken and involved dropping more than 300 tonnes of poisoned bait (**pictured**) over 1,087 square kilometres of the island — by hand and by helicopter. The South Georgia Heritage Trust announced on 8 May that a two-year survey of the island with sniffer dogs had found no remaining trace of rodents. Mice and rats first arrived on South Georgia, a British Overseas Territory, on vessels in the late 1700s, and have had a devastating effect on endemic bird species. Some of these birds are now showing signs of recovery, the trust said.

TONY MARTIN/SGHT

Grizzly proposal

US officials are considering removing Endangered Species Act protections for grizzly bears in the Northern Continental Divide Ecosystem, the US Fish and Wildlife Service announced on 9 May. The ecosystem covers more than 30,000 square kilometres of the northern Rocky Mountains, and includes Glacier National Park. There were an estimated 942 grizzly bears (*Ursus arctos horribilis*) in this area in 2011. Removing federal protections would turn management of the animals over to the state of Montana, where the bears live. The announcement comes almost a year after the agency removed grizzlies in the greater Yellowstone ecosystem from the list of endangered species.

PEOPLE

UK society fellows

Britain's Royal Society named 50 eminent scientists as its newest fellows on 9 May. Among them are inventor Elon Musk, theoretical physicist and broadcaster Jim Al-Khalili, plant scientist Cathie Martin and artificial-intelligence researcher Demis Hassabis. David Willetts, who was the UK government's universities and science minister from 2010 to 2014, was named an honorary fellow for his consistent championing of science. Of the new fellows, 12 are women. Scientists, engineers and technologists are elected to the society on the basis of their exceptional contributions to science.

Bacteria pioneer

Microbiologist Stanley Falkow has died at the age of 84, Stanford University in California said on 10 May. Falkow, an emeritus professor at Stanford, was known for his research on how bacteria cause human disease and how antibiotic resistance spreads. One of his biggest discoveries — made in the 1970s — was that bacteria can transfer circular pieces of DNA known as plasmids to each other. Falkow also advocated for US federal regulators to ban the use of antibiotics in animal feed. A winner of the 2015 National Medal of Science, he trained more than 100 students and postdocs during his career. Falkow died in his California home on 5 May.

BUSINESS

Listing called off

Academic publisher Springer Nature has postponed its plans to float on the Frankfurt stock exchange in Germany. The company said on 8 May that market conditions had delayed plans to raise €1.2 billion (US$1.4 billion) by selling new shares in an initial public offering. The firm had intended to list on the stock exchange on 9 May. Springer Nature, which publishes this journal and around 3,000 other titles, was formed in 2015 in the merger between London-based Macmillan Science and Education and Berlin-based Springer Science+Business Media. A large portion of the proceeds from the planned listing were to be used to cut the company's debt. (*Nature*'s news team is editorially independent of its publisher.)

POLICY

Clinic crackdown

The US Food and Drug Administration (FDA) announced on 9 May that it is seeking permanent injunctions to halt the sale of unapproved stem-cell treatments by two clinics. The agency says that one of the companies — US Stem Cell Clinic of Sunrise, Florida — has continued to market unapproved products to patients, despite receiving a warning letter from the FDA last August. The FDA says that the second company used live-smallpox vaccines in treatments given to people with cancer. Federal marshals seized vials of the vaccine in 2017 after the agency inspected clinics run by the California Stem Cell Treatment Center, an affiliate of the Cell Surgical Network Corporation, based in Rancho Mirage and Beverly Hills, California.

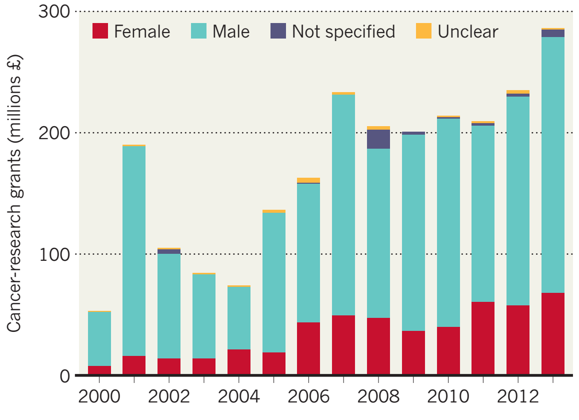
➲ NATURE.COM
For daily news updates see:
www.nature.com/news

TREND WATCH

Nearly 70% of UK cancer-research grants awarded over a 13-year period went to male researchers, an analysis finds. A team analysed the details of about 4,200 grants given out by public and philanthropic funders to UK institutions from 2000 to 2013, together totalling £2.33 billion (US$3.2 billion). Female investigators received just 31% of the grants, and the median award value was 1.3 times greater for men. The discrepancy persisted across types and stages of research, and across funders.

CANCER RESEARCH'S GENDER GAP

Most cancer-research grants awarded to UK researchers go to male scientists.



SOURCE: C. D. ZHOU *ET AL. BMJ OPEN* **8**, E018625 (2018).

NEWS IN FOCUS

LAB LIFE Prominent cell biologist attempts to revive career **p.288**

CONSERVATION How Europe is demolishing dams to restore its fractured rivers **p.290**

BIBLIOMETRICS The most-cited journal articles on Wikipedia **p.291**

WORK LIFE A *Nature* survey finds leadership problems in research **p.294**

ABEDIN TAHERKENAREH/EPA/REX/SHUTTERSTOCK

Inspectors from the International Atomic Energy Agency tour an Iranian facility in 2007.

POLICY

US exit from Iran nuclear deal endangers research

International collaborations could end in wake of US President Donald Trump's decision.

BY JEFF TOLLEFSON

The United States will no longer be part of the Iran nuclear deal, delivering a major blow to efforts to establish scientific collaborations between the two countries. Researchers say that the decision, announced on 8 May by US President Donald Trump, will make a bad situation worse.

Under the 2015 Joint Comprehensive Plan of Action (JCPOA), Iran agreed to scale back its nuclear programme and allow international inspections of its facilities in exchange for the removal of economic sanctions imposed by the United States, the European Union, Britain, Russia and China. At the time, many researchers saw the agreement as an opportunity to bolster Iranian science and to expand international collaborations.

But those plans have encountered roadblocks since the 2015 deal. For example, when Trump took office last year, long-standing efforts to establish scientific exchanges between Iran and the United States came to a halt. And workshops organized by the US National Academies of Sciences, Engineering, and Medicine (NASEM) between 2010 and 2017 — meant to encourage collaborations in diverse fields including solar energy and water resource ▶

management — stopped after the Trump administration raised questions about Iran and the nuclear deal, says Glenn Schweitzer, who spearheaded the NASEM work in Washington DC.

"We were all full of enthusiasm when the agreement was signed, but unfortunately things went in the opposite direction," says Soroosh Sorooshian, an Iranian–American hydrologist at the University of California, Irvine. He was one of hundreds of scientists who participated in the NASEM workshops. "God knows what happens next."

MOTHBALLED

Iranian scientists have expanded collaborations with their European counterparts in areas such as nuclear safety and security, but similar work has failed to take root in the United States.

That is in part because some US sanctions remained in place in spite of the nuclear agreement, and because US researchers often need a licence from the US Department of the Treasury to collaborate with government scientists in Iran, says Matthew Bunn, who studies nuclear non-proliferation issues at Harvard University in Cambridge, Massachusetts.

Bunn is seeking such a licence to initiate a dialogue with leading nuclear scientists in Iran, with the ultimate goal of steering the country towards a safe and secure nuclear-energy programme. Trump's decision could detract from efforts to advance meaningful scientific cooperation, Bunn says, in addition to emboldening Iranian hardliners who would like to see the country become a nuclear power.

"There won't be a lot of enthusiasm on the Iranian side for dialogues with Americans."

"I need to rethink what I had been planning," he says. "There won't be a lot of enthusiasm on the Iranian side for dialogues with Americans such as myself."

Other research collaborations that could be in jeopardy include work at Fordow, an underground nuclear facility near Qom in northern Iran. As part of the JCPOA, Iran agreed to halt uranium enrichment at the facility. The country planned to pursue particle-physics research there, and to use the facility to produce medical isotopes. Russian scientists had been working with Iran on experiments to advance Iran's medical isotope production, says Scott Kemp, who heads the Laboratory for Nuclear Security and Policy at the Massachusetts Institute of Technology in Cambridge.

"I think that work gets mothballed, at least at the outset," Kemp says. And if the agreement collapses entirely and Iran walks away, he says that the country would scrap the effort altogether "and go back to making enriched uranium".

Sorooshian says the only good news is that the number of Iranian students entering US universities has increased in the past few years, which will help to build relationships between the two countries in the decades to come.

But for now, he says that the outlook for scientific cooperation between the two countries looks grim. "Everybody is concerned." ■

PEOPLE

Sacked Japanese biologist to retrain at Crick Institute

Yoshinori Watanabe hopes to revive his career with help from his former mentor.

BY DAVID CYRANOSKI

Prominent cell biologist Yoshinori Watanabe, who was dismissed by the University of Tokyo last month, is attempting to put his past behind him by embarking on an intensive retraining programme with Nobel prizewinner Paul Nurse in London. The university dismissed Watanabe after an investigation concluded that he had committed scientific misconduct.

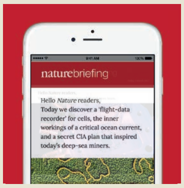
Watanabe, who has done groundbreaking work in chromosome biology and has a string of impressive scientific achievements to his name, arrived at Nurse's laboratory on 16 April. Watanabe says the programme will focus on data acquisition and presentation, and also involve experiments. "After that period of retraining, I hope that I will be able to find somewhere to continue my research career," he says. Watanabe told *Nature* that he made mistakes in scientific papers, but that he did not intend to deceive and that he thinks these errors do not amount to serious misconduct.

Programmes to retrain errant scientists are rare. A rehabilitation initiative run by ethicist James DuBois at Washington University in St Louis, Missouri, with support from the US Office of Research Integrity, trained 61 researchers between 2013 and 2017. Participants who are referred to the programme have generally made careless mistakes, failed to provide adequate oversight, or not complied with policies on the treatment of human research participants, animal welfare or the declaration of conflicts of interests. But few of the rehabilitation participants have been accused of manipulating data, as Watanabe was.

Nurse, who mentored Watanabe when he was a postdoctoral researcher in the 1990s, thinks that the biologist deserves the

MORE ONLINE

NATURE BRIEFING



Save time — get *Nature*'s daily newsletter delivered straight to your inbox go.nature.com/savetime

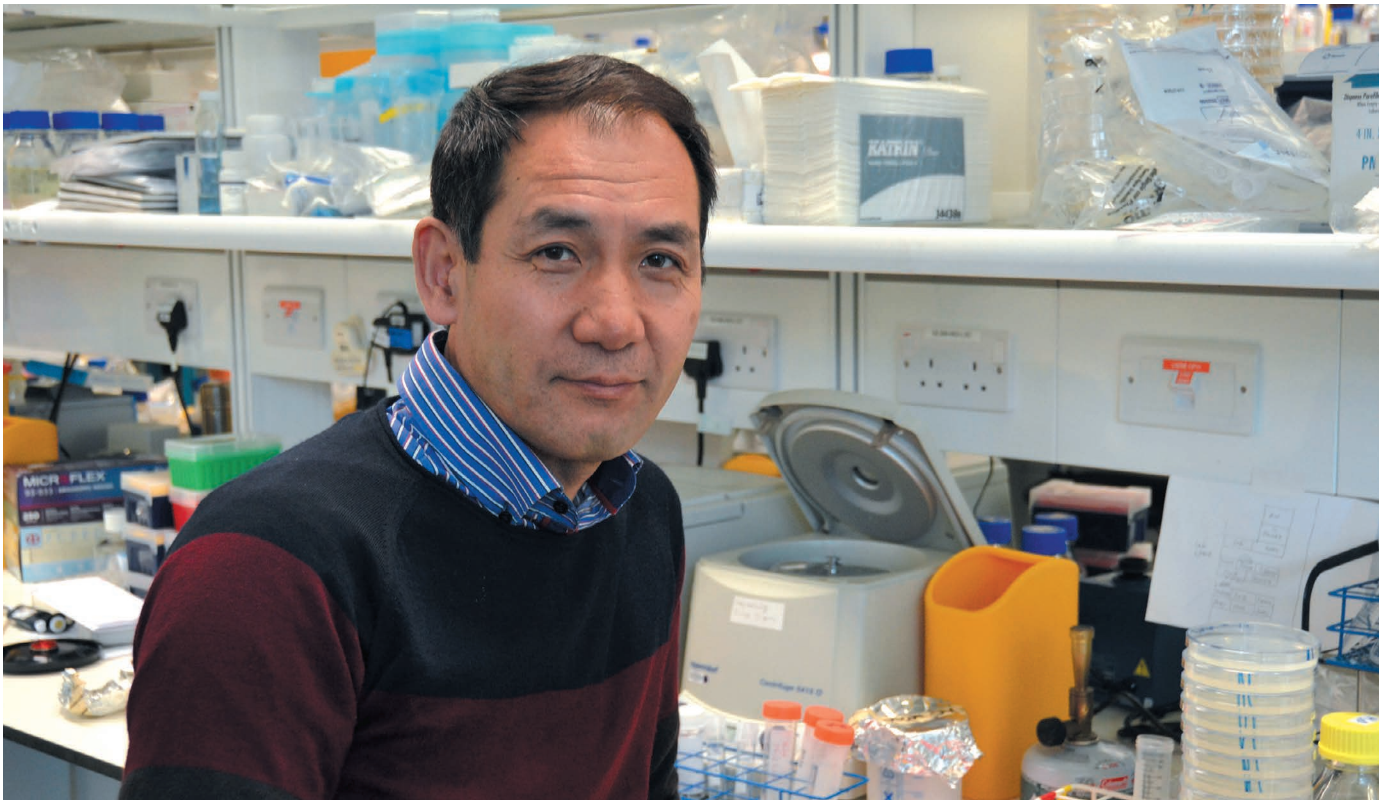
TOP NEWS

- AI recreates activity patterns that brain cells use in navigation go.nature.com/2l1ryrz
- Why it's hard to prove gender discrimination go.nature.com/2lhcuxs
- Butterflies caught in an evolutionary trap go.nature.com/2g0wyou

NATURE PODCAST

Probing the proton; research misconduct; and making sense of mystery genes nature.com/nature/podcast

MATTHEW BUTT/CRICK INST.



Yoshinori Watanabe, who has done groundbreaking work in chromosome biology, will attend a retraining programme in London.

opportunity to redeem himself. "The research community and institutions need to think more about how to handle rehabilitation in cases like this," says Nurse, a cell biologist and director of the Francis Crick Institute in London. Nurse declined to comment further on the retraining. A spokesperson for the institute says that "it isn't part of a formal approach being taken by the Crick. It's a one-off situation — Paul Nurse agreed to give Dr Watanabe an opportunity for retraining."

SOUND RETRAINING

Frank Uhlmann, a cell-division researcher at the Crick, has been advising on Watanabe's retraining. He suggests that Watanabe's experiments at the Crick include the use of automated image-acquisition and analysis tools, and a double-blind laboratory set-up that conceals from the experimenter which data are the tests and which the controls. Watanabe told *Nature* he intends to do both.

Bill Earnshaw, a cell biologist at the University of Edinburgh, UK, says he is happy that Nurse is giving Watanabe this opportunity. "But whether this will succeed, only time will tell — as I know of no precedent for this. I am supportive and cautiously optimistic," he says.

Watanabe says he is excited about the opportunity to retrain at the Crick, but thinks it will be a challenge to secure a new research position.

After a career spanning more than three decades, Watanabe's experiments continue to form the bedrock of scientists' understanding of how proteins drive the separation of chromosomes when cells divide. "His major findings have been correct and influential," says Arshad Desai, a cell biologist at the University of California, San Diego.

But last August, the University of Tokyo announced that five of Watanabe's papers contained manipulated images and improperly merged data sets that amounted to scientific misconduct. One of those papers has since been retracted and two have been corrected. Two others have corrections under consideration, according to Watanabe. Another university investigation into nine other papers found no evidence of misconduct.

Watanabe says that the university's investigation made him aware of "issues concerning contrast in pictures and checking original imaging files". He says, however, that he did not intend to deceive and that the issues did not affect the main conclusions of the papers. Although Watanabe resigned from the university on 28 February this year, his official dismissal from the university means he will not receive a pension.

"Should Paul urge Watanabe's rehabilitation, then I think there's some hope Watanabe will be able to return to science."

SECOND CHANCE

Many of Watanabe's colleagues think he deserves an opportunity to make amends. Julia Cooper, a molecular biologist at the US National Cancer Institute in Bethesda, Maryland, says that data manipulation is never acceptable. But she thinks the sanctions were too harsh and incommensurate with the degree of wrongdoing. "Yoshinori absolutely deserves a second chance," she says.

SENSITIVE SITUATION

Several people contacted by *Nature* did not want to be quoted, citing either the sensitive nature of the situation or its complexity. But most have faith in Nurse's judgement. "Should Paul urge Watanabe's rehabilitation, then I think there's some hope Watanabe will be able to return to science," one researcher said. Another expects that Watanabe's retraining will be successful.

Despite helping to design Watanabe's rehabilitation programme, Uhlmann is unsure whether it will make a difference. He commends Watanabe's willingness to engage with his retraining, but says "we will only know at the end of it whether his heart is where his mouth is".

Watanabe emphasizes that his willingness to embark on the training and acknowledgement that he made errors is evidence that he will change his ways.

Silke Hauf, who was one of Watanabe's postdoctoral researcher and is now a cell biologist at Virginia Tech in Blacksburg, says that Watanabe deserves another chance, but is likely to struggle to regain his career even after retraining. "No institution wants to give the appearance of fostering misconduct," she says, "but allowing people to rectify mistakes must be part of a scientific culture." ■

FUNDING

Australian science boost

Government pushes for long-term investment and more cash.

BY NICKY PHILLIPS

Research facilities and medicine were among the winners for science in Australia's 2018–19 national budget, which was proposed on 8 May.

Prime Minister Malcolm Turnbull's centre-right Liberal government will push to invest almost Aus$1.9 billion (US$1.4 billion) over the next 12 years in shared research infrastructure — such as microscopes, supercomputers, a marine observing system, and telescopes used in myriad disciplines, from nanotechnology to oceanography. The new money is in addition to the $2.2 billion-plus that the government has already committed to spend on research facilities and programmes over 10 years, as announced in December 2015.

"We're pretty excited about that," says Kylie Walker, chief executive of Science & Technology Australia in Canberra, an umbrella organization of scientific societies. The final budget will require the Senate's approval.

Medical research will receive an increase of $1.3 billion, to be spent on various programmes and initiatives over 10 years that aim to improve health and boost the medical industry, including $500 million for a genomics and precision-medicine initiative. The government currently spends about $1.2 billion a year on medical research, according to the Australian Academy of Science.

SPACE CASH

The budget also sets aside $4.5 million over 4 years to encourage more women to study and work in science, technology, mathematics and engineering, and $26 million in seed funding to establish a national space agency. Andrew Dempster, director of the Australian Centre for Space Engineering in Sydney, says that, as well as money, the national space agency needs bipartisan support from both sides of government: "Space programmes are longer than the electoral cycle."

> ***"Space programmes are longer than the electoral cycle."***

Rosalind Dubs, board director of the Australian Academy of Technology and Engineering in Melbourne, told *Nature* that a national space agency will ensure Australia has clout when interacting with other countries' space agencies, and will guarantee Australia's continued access to Earth-observation and global-positioning satellites.

Researchers have been lobbying for several years for long-term investment in Australia's shared research facilities, under the National Collaborative Research Infrastructure Strategy (NCRIS). "Research infrastructure has been funded on a year-to-year basis for a long time, which made it very vulnerable," Walker says. A political stalemate in 2015 between the centre-right Liberal government of then-Prime Minister Tony Abbott and the opposition Labor Party meant that many NCRIS facilities nearly ran out of money to keep their sites open and pay staff.

Beyond specifying cash for new supercomputers and the Australian Animal Health Laboratory in Geelong, the budget did not state which infrastructure projects will receive funding. But Walker says that the extra money will allow facilities to attract and retain a skilled workforce, and invest in capital works to maintain and upgrade equipment.

The government has also allocated almost $500 million to help heal the ailing Great Barrier Reef. That funding, announced on 29 April, includes $444 million for the Great Barrier Reef Foundation in 2017–18. Although some researchers welcomed the proposal, others say that it will not address the biggest threat to the reef's health — global warming. ■

ECOLOGY

Dam removal restores rivers

Huge European demolition projects offer hope for fragmented ecosystems.

BY QUIRIN SCHIERMEIER

The Yecla de Yeltes Dam in western Spain supplied drinking water to local communities for half a century, until newer projects rendered it obsolete. Its demolition this month is the biggest dam-removal project in the European Union so far — and is being hailed by ecologists as a milestone for river-restoration efforts on the continent.

Such efforts are ramping up in many European countries — although others, notably those in the Balkan Peninsula, are on a dam-building spree. An initiative has begun to take the first continent-wide census of all dams. And although dam removal is generally welcomed by most scientists, some call for more research into potential ill effects.

Hundreds of thousands of dams and weirs, most small and many no longer in use, fragment Europe's rivers. The structures, some of them thousands of years old, have provided irrigation, energy and other benefits. But their presence also threatens the habitats of endemic fish and wildlife.

"Dams alter the natural characteristics of a river system," says Jeroen van Herk, a project manager for Dam Removal Europe, a group that promotes river restoration. "Long stretches of rivers, which once flowed freely from source to outlet, become a series of pools, hindering migrating fish from reaching spawning grounds in the upper reaches."

The Yecla de Yeltes is on the Huebra River, a 122-kilometre-long tributary of the Duero, which is one of the Iberian Peninsula's main rivers. Ecologists suspect that the 22-metre-tall dam, built in 1958, is partly responsible for the observed decline of a small freshwater fish called the sarda (*Achondrostoma salmantinum*), along with that of other endemic species, including otters and black storks (*Ciconia nigra*), which were once abundant in the area. Scientists in Spain are set to monitor whether the animals come back after the dam is removed.

Across much of Europe, rivers unfettered by artificial barriers are exceedingly rare. However, over the past 20–25 years, at least 5,000 small dams, weirs and culverts have been removed from rivers in France, Sweden, Finland, Spain and the United Kingdom, according to Dam Removal Europe. (There are few reliable records from other European countries.)

Dam removal gained momentum after the EU adopted the Water Framework Directive in 2000, legislation that requires member states to improve the ecological protection of rivers and lakes. But as yet, only about half of rivers in the EU meet its environmental objectives, says Wouter van de Bund, an aquatic ecologist at the European Commission's Joint Research

DAM REMOVAL EUROPE

The Yecla de Yeltes Dam in Spain is being demolished in the largest project of its kind ever in Europe.

Centre (JRC) in Ispra, Italy.

In the United States, about 1,200 barriers have been dismantled in recent decades, with generally positive effects on local ecosystems, says Laura Wildman, a fisheries engineer at eco-consultants Princeton Hydro in South Glastonbury, Connecticut.

But restoration projects need to be monitored for negative effects, too, experts say. Decommissioning existing river barriers might mobilize toxic sediment, or affect buildings or bridges downstream. And existing dams could help to prevent the spread of invasive species such as the North American signal crayfish (*Pacifastacus leniusculu*) or the Asian topmouth gudgeon (*Pseudorasbora parva*). There are also historic dams, such as the Roman-built, 22-metre-high Prosperina Dam near Mérida in Spain, which need to be preserved as cultural heritage.

Dams were built with little regard for the impacts they might have on ecosystems, says Carlos Garcia de Leaniz, an ecologist at Swansea University, UK, who coordinates the US$6.2-million, EU-funded Adaptive Management of Barriers in European Rivers (AMBER) project. "We must not make the same mistake when dams are being removed."

In collaboration with the JRC, AMBER is managing an exercise to map the location of all registered dams and weirs in 38 countries across the continent, including some that aren't EU members. The project database currently holds information on 230,000 river barriers in 13 EU countries. Drawing from nine case studies, the project also aims to develop tools to help water authorities assess the costs, benefits and damage potential of dam-removal projects.

Inventories are important for planners and policymakers to understand the scale of issues caused by river fragmentation, says Wildman.

A number of small dams in the Netherlands, Denmark and Spain are scheduled for removal later this year. And starting in 2019, French scientists plan to systematically monitor the impacts of a removal project even larger than Yecla de Yeltes: the demolition of two hydropower dams in the Sélune Valley in Normandy, one 35 metres tall and the other 15 metres.

But while old barriers are being removed, new dams are built elsewhere. Some 2,800 hydropower plants are currently being planned across the Balkans — a threat, says van de Bund, to many of the continent's last untouched rivers. ■

BIBLIOMETRICS

Wikipedia's top-cited scholarly articles — revealed

Gene studies dominate lists of DOI publications referenced highly in the encyclopaedia.

BY GIORGIA GUGLIELMI

The most-cited journal articles on Wikipedia include papers on the names of lunar craters and the DNA sequences of genes — and many of these works are referenced more times in the online encyclopaedia than they are in the scientific literature.

"It is pretty incredible that almost all the highly cited articles are science articles," says Matt Miller, a data scientist and librarian based in New York City. Miller analysed citation data released in March by the Wikimedia Foundation, the non-profit organization in San Francisco, California, that runs Wikipedia. The data set — which contains some 15.7 million records — shows how many times sources with formal identifiers such as ISBNs (international standard book numbers) and DOIs (digital object identifiers) are referenced across all of Wikipedia's nearly 300 language editions. ▶

WIKIPEDIA’S TOP JOURNAL ARTICLES

The five most-cited DOIs across language editions

2,830,341 citations ‘Updated world map of the Köppen–Geiger climate classification’ (2007)
21,350 citations ‘Prediction of hydrophobic (lipophilic) properties of small organic molecules using fragmental methods’ (1998)
20,247 citations ‘The status, quality, and expansion of the NIH full-length cDNA project: The Mammalian Gene Collection (MGC)’ (2004)
5,937 citations ‘Generation and initial analysis of more than 15,000 full-length human and mouse cDNA sequences’ (2002)
5,854 citations ‘The Asiago supernova catalogue — 10 years after’ (1999)

▶ Most publications cited using identifiers on Wikipedia are books, but Miller looked at the numbers for publications with DOIs — the most widely used identifier for journal articles — on English-language Wikipedia. His data set contains 1.2 million citations that used DOIs, referencing more than 835,000 unique articles.

The most-referenced paper, with 4,702 citations across English Wikipedia, is a 2002 collection of more than 15,000 sequences of human and mouse genes. The Wikipedia pages that reference the study are mostly entries about single genes or proteins. In all, five articles in the top ten are about DNA catalogues. (Wikimedia’s original post notes: “Unsurprisingly, Wikipedians love reference works.”)

Four astronomy articles feature in the top ten, including studies on star distances, asteroids and the names of lunar craters (the last in a 1971 paper that has just 16 citations in the scientific literature, according to Google Scholar).

A separate analysis of the wider Wikimedia data by Ross Mounce, who directs open-access programmes at the London-based philanthropic foundation Arcadia Fund, reveals the most-cited DOI articles across all of the encyclopaedia’s language editions (see ‘Wikipedia’s top journal articles’; for the full lists, see go.nature.com/2gfnfui). Six articles in the top ten are the same as for the English-language version, but the first entry is notably different: a 2007 paper on the global climate, which has 2.8 million citations — but only 169 on English Wikipedia.

The climate study is so heavily referenced because millions of its citations come from pages created by an automated computer program. The bot, developed by physicist Sverker Johansson at Dalarna University in Falun, Sweden, had produced nearly 3 million articles as of July 2014, according to Wikipedia. One-third of the articles, many of them about geographical locations, are in Swedish and the rest are in Cebuano and Waray, two languages spoken in the Philippines.

“One of the most interesting things is that this information is available at all,” says John Chodacki, director of the University of California Curation Center, who is based in Berkeley. Analysing and comparing citation data across scholarly papers has historically been possible only using paywalled services. ■

Additional reporting by Richard Van Noorden.

CORRECTION

The News Feature ‘After the violence’ (*Nature* **557**, 19–24; 2018) erroneously stated that Colombia’s reincorporation efforts would cost 129.5 billion pesos (US$46 million). Actually, they would cost 129.5 trillion pesos (US$46 billion).

SPECIAL ISSUE

HOW TO GROW A HEALTHY LAB

A SPECIAL ISSUE ON HOW TO MAKE LAB GROUPS PRODUCTIVE, RIGOROUS AND HAPPY.

If you find a bad apple, check the barrel. Research-integrity specialists say that focusing too much on individual bad actors deflects attention from the environments that promote bad behaviour. The idea applies just as much to researchers who are unproductive, frustrated or unhappy — they could be indicative of deeper problems.

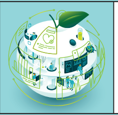
A healthy research environment is fundamental to good science: it helps people to produce their best work, and feel satisfied in doing so. But the matter is rarely discussed. That's partly because a lab's 'health' is complex and difficult to assess — the product of a mix of factors, such as inclusivity, communication, career pressures and training. And academics often feel ill-equipped to tackle these matters.

In this issue, we explore how the working environment shapes research quality and morale — and what people can do to strengthen the research enterprise. A survey of more than 3,000 researchers on page 294 reveals that lab heads view their labs' practices more positively than do trainees. Roughly 40% of junior scientists say that their labs sometimes cut corners to achieve a desired end.

Some institutions are taking steps to improve lab health. In a World View on page 281, Catherine Winchester describes how her institute created the post of research-integrity officer to champion best practices. In a Comment on page 299, researchers at the University of California, Berkeley, explore how clear expectations in the university's College of Chemistry might prevent graduate students from falling through the cracks. And on page 297, experts at the National Center for Professional & Research Ethics at the University of Illinois at Urbana–Champaign outline pitfalls that can plunge researchers into scandals.

Advice is at hand. On page 302, scientists share their top tips for improving research culture. And on page 457, five leadership specialists share lessons from their respective fields.

Improving the culture of a lab group or research institution is no small task. But both institutions and individuals can take concrete steps that get to the core of the matter — and everyone can benefit from that. ■



LAB HEALTH
A *Nature* special issue
nature.com/collections/labhealth

ILLUSTRATION BY MARCO GORAN ROMANO

ILLUSTRATION BY MARCO GORAN ROMANO

LEADERSHIP PROBLEMS IN THE LAB

A survey of 3,200 scientists reveals the tensions bubbling in research groups around the world.

BY RICHARD VAN NOORDEN

Scientists pride themselves on being keen observers, but many seem to have trouble spotting the problems right under their noses. Those who run labs have a much rosier picture of the dynamics in their research groups than do many staff members working in the trenches, according to a *Nature* survey of more than 3,200 scientists. The results suggest that a lack of training in lab and personnel management is one of the strongest contributors to an unhealthy lab culture.

"The communication of experiences between senior and junior researchers is dismal," says Gary McDowell, executive director of the Future of Research, a non-profit organization in San Francisco, California, that advocates on behalf of young scientists. "They live almost in separate worlds."

Concern over the integrity of the research enterprise has been mounting for years, and has triggered high-level studies by the US National Academies of Sciences, Engineering, and Medicine and the US Office of Research Integrity, among others. They have been looking to understand training gaps and the ever-increasing pressure on scientists to secure grants, publish papers and earn promotions.

Last year, *Nature* convened 16 meetings and workshops in universities across Europe and the United States to explore the state of lab health, pressures on individual groups and how best to tackle them. Scientists shared what they liked and loathed about their workplace, from navigating interpersonal relationships to enforcing and encouraging best practices. *Nature*'s survey grew from these discussions, in an effort to back up such anecdotes with data. It is the largest publicly reported analysis of its kind.

The encouraging news is that morale is reasonably high. For the most part, scientists around the world view their groups as healthy — calling them 'friendly', 'collaborative' and 'supportive'. But signs of stress bubble underneath the surface: around one in five respondents in more junior positions (that is, those who don't lead the group, such as graduate students and postdoctoral fellows) were negative about their labs, describing them as 'stressful', 'tense' and 'toxic' (see 'Words matter').

Scientists who took the survey said they wanted more principal investigators (PIs) to take training courses, and suggested that PIs ask for feedback from their lab groups more regularly. And slightly more than half of non-PIs said they had often or occasionally felt pressured to produce a particular result in the past year. "That's very alarming," says Jeffrey Flier, a physician-researcher at Harvard Medical School in Boston, Massachusetts, and dean of the faculty of medicine until 2016.

Nicholas Steneck, who studies research integrity at the University of Michigan in Ann Arbor, cautions that the survey, which was sent to more than 250,000 *Nature* readers and advertised on the magazine's website, reflects a limited

LAB HEALTH
A *Nature* special issue
nature.com/collections/labhealth

sample of scientists and might over-represent those who are interested in complaining. But he argues that it still shows an unacceptable level of tolerance for bad behaviour. "I am not sure I see any good news in the survey, just different degrees of bad news," he says.

SPLIT PERSPECTIVE

The senior researchers who responded — 655 lab heads — were hugely positive about their workplace practices. More than 90% said that they consistently reviewed their research group's experimental designs and could summarize which projects every member of their group was working on. They were equally confident about their availability to discuss staff members' experiments or career development; that they valued negative results; and that they gave their scientists freedom to explore interesting findings that were not necessarily core to the lab's research activities.

The 2,632 non-PIs sampled were much less optimistic (see 'Perception gap'). About 80% agreed their PIs could summarize each group member's project; around 70% that their PI was readily available for discussions and let them explore non-core results. Just two-thirds said that their PI consistently reviewed the design of experiments or valued negative results.

A particular disconnect came in attitudes towards checking raw data. Just 57% of lab members said their PIs consistently checked raw data — but 90% of PIs said they did. Given the time crunch involved in juggling staff management with applying for grants and other duties, "I doubt if such PIs will check raw data with the consistency they claim," says Angela Goh, a computer scientist who retired this year as associate provost at Nanyang Technological University in Singapore. The gap is striking, and it might be that PIs and other researchers have different definitions of 'raw data', says C. K. Gunsalus, who studies research misconduct at the University of Illinois at Urbana-Champaign.

Radovan Šebesta, who leads an organic chemistry group at Comenius University in Bratislava and took the survey, was surprised by the findings and a little concerned about the problems identified by junior researchers: "Is what I do perceived in the same way by my group?" Šebesta is thinking of asking his lab members how they would have answered the questions.

Most of the survey respondents chose to remain anonymous, so it is not possible to match PIs and non-PIs to the same labs. But the two groups did not show significant geographical differences. And the findings dovetail with established social-psychology results about the dynamics of organizations, says Gunsalus. "The more power you have, the less attuned you are to how actions are perceived at the lower levels," she says.

The same perception gap was evident in other parts of the survey. Almost 90% of PIs felt that their lab or group members were clear about what was expected of them, and two-thirds said the group 'never' or 'rarely' condones research practices that cut corners — such as valuing speed over quality, or fundability over accuracy. Only two-thirds of non-PIs said that they personally were clear about what their PI expected of them, and only 43% felt their group never or rarely cut corners.

Although these opinions and attitudes might not reflect a reality of wrongdoing in labs, Steneck and others see them as warning signs. Previous research on individual organizations has shown that scientists who perceive their lab climates to be favourable are less likely to say they partake in practices that would be construed as misconduct, such as plagiarism, falsification and fraud.

THE DISCONTENTED

One goal of the survey was to understand how lab culture can promote or hinder research. Brian Martinson, who studies research integrity with the HealthPartners Institute in Bloomington, Minnesota, says that *Nature*'s survey found a moderate correlation between those who described their lab in negative terms and those who said this atmosphere seriously hindered the lab's ability to produce good-quality research. (It was also clear that some respondents had misread the latter question, describing their lab's culture in glowing terms but saying that it hindered their ability to do research. Follow-up contact clarified the confusion in some cases.)

To get a sense of researchers who felt truly dissatisfied with their lab culture, *Nature* identified a group of scientists who were consistently negative about their experiences. They used words such as 'abusive', 'oppressive' and

PERCEPTION GAP

Principal investigators (PIs) and researchers in more junior positions have different views on how involved lab leaders are in the work of the group.

Totals may not add to 100% due to rounding.

WORDS MATTER

The most common words respondents used to describe their labs were mostly positive, but a subset of generally discontented researchers (about 14% of the non-principal investigators) chose more negative terms.

Word rank	Tally
1. FRIENDLY	595
2. COLLABORATIVE	480
3. SUPPORTIVE	342
4. COMPETITIVE	207

Word rank	Tally
1. COMPETITIVE	45
2. STRESSFUL	35
3. TENSE	19
4. TOXIC	16

'hostile', and reported that the bad vibes in their lab harmed their work. These discontented researchers made up around 14% of the non-PIs — 376 scientists. Although they do not represent all the unhappy respondents, they reflect a group that clearly and consistently voiced frustration.

Many of them seemed to be dissatisfied with lab leadership; their perceptions of PI behaviour were relatively negative. Just 20% felt their lab's culture never or rarely condoned corner-cutting, and only 38% felt their PI was almost always available for discussions. And 70% of them said that in the past 12 months they had 'often' or 'occasionally' felt pressured to produce a particular result, compared with slightly more than half of all non-PI respondents. It can't be known whether this group is genuinely experiencing a much worse lab culture than other respondents, Gunsalus notes. "But there are some horrible environments out there, for which institutions should be held accountable," she says.

HOW TO HELP LABS DO BETTER SCIENCE

In one of the survey's starkest findings, two-thirds of researchers who head laboratories said they had not had training in managing people or running a lab in the past year — and the majority of those said they wanted some (see 'Training gap'). "One of the biggest surprises of my PI career is how difficult the management part is," says Šebesta. "I would love to have had some training." And of the PIs who had received training, five-sixths found it useful.

Many institutions, including the Howard Hughes Medical Institute in the United States and the European Molecular Biology Organization, have set up highly subscribed courses in mentoring and management. But most don't mandate such training, says Flier. That might be partly out of reticence to force it on academics, he says — and because "institutions don't rate it that highly".

The results suggest that training is sorely needed, says Michael Mumford, a psychologist at the University of Oklahoma in Norman. "We take people with no management or leadership training and tell them to run a team of 3–20 people and assume they should just know how to interact and how to manage others," he says.

"All my friends who have recently become PIs wish they had some sort of training," says one US biology postdoc who took the survey. "They have been thrown into the wilderness without learning any of the skills needed to survive."

Asked to choose from a list of options for how their department head or institution could help their lab to produce higher-quality research, more than 60% of the PIs said they wanted more support for mentoring and managing (see 'Factors to fix'). That desire came second only to a request for more resources for administrative tasks.

More than 40% of the non-PIs also said their lab could do better science if PIs received more training in mentoring and management. Among the least happy, more than 70% indicated that this was their main desire. "I believe my institution has courses for PIs, but for some reason, they are optional. I think they should be a requirement," wrote one microbiology graduate student in Germany.

Asked whether any other factors might help, more than half of lab heads mentioned funding — a common refrain in discussions about limits on research. "The competition for funding, as well as the short-term nature of the funding one gets, creates an atmosphere where scientific reproducibility counts far less than getting the next grant," wrote one US lab head.

The survey should be a signal that even if PIs think their lab has a healthy research culture, it is important to revisit communication and goals, and to seek feedback from those they manage, says Gunsalus. "Even if your own institution doesn't provide it, there are mentoring resources out there," she says. In the United States, she points people to the National Research Mentoring Network and a collection of articles on leadership from the National Center for Professional & Research Ethics at the University of Illinois at Urbana-Champaign.

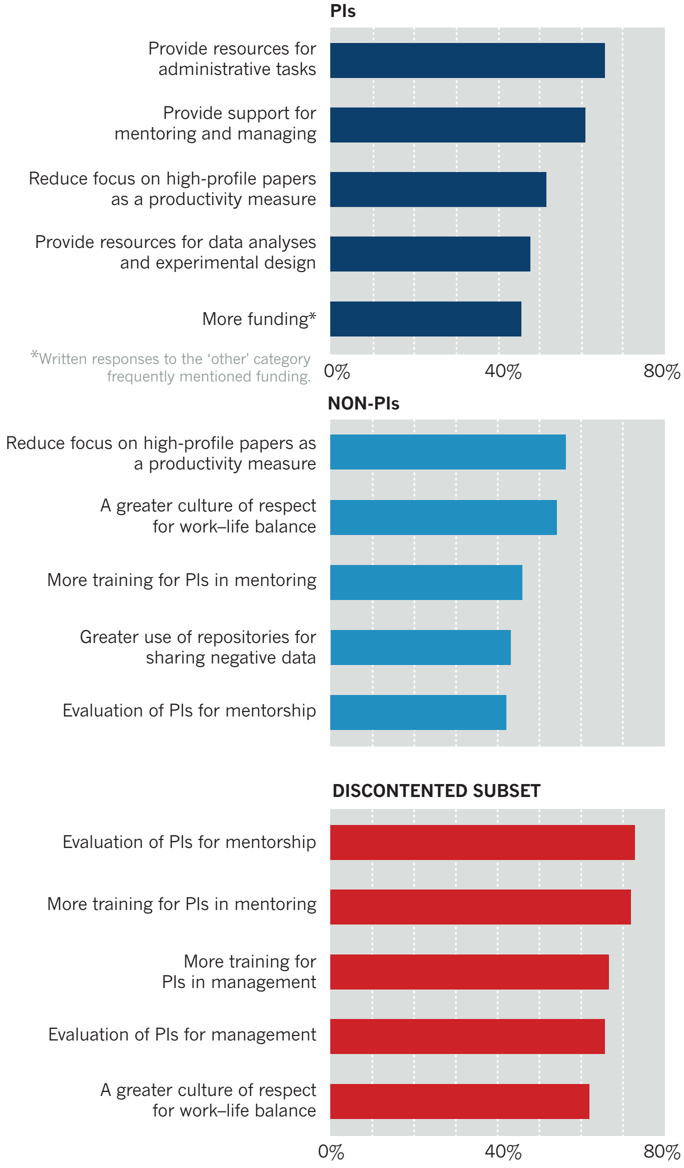
"Institutions need to step up their provision of effective resources for lab leaders and professional development for emerging researchers," she says. "They should 'own' the research environments they are providing." ■

Richard Van Noorden *is a news features editor with* Nature *in London.*

For a copy of the survey and responses, visit go.nature.com/2rzlo91

FACTORS TO FIX

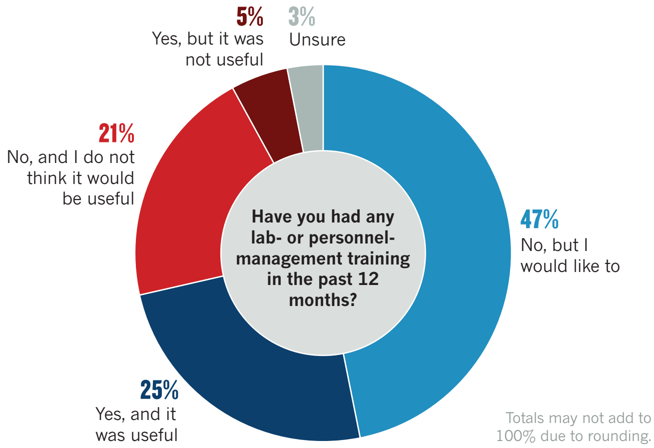
When asked what would improve their labs, principal investigators (PIs) and non-PIs have different priorities. A subset of scientists who are consistently discontented with their working situation focus on factors that would improve leadership.



*Written responses to the 'other' category frequently mentioned funding.

TRAINING GAP

Nearly half of the principal investigators in our survey want training in managing people or running a lab but haven't been able to get it recently.



Totals may not add to 100% due to rounding.

COMMENT

LAB HEALTH Clear expectations enable women and minorities to flourish **p.299**

LAB HEALTH Five experts each pick one change to help research groups thrive **p.302**

PSYCHIATRY Ramifications of Hans Asperger's Nazi collusion **p.305**

BIOTECHNOLOGY Why did so many believe the Theranos story for so long? **p.306**

ILLUSTRATION BY DAVID PARKINS

Nine pitfalls of research misconduct

Academic leaders must audit departments for flaws and strengths, then tailor practices to build good behaviour, say **C. K. Gunsalus** and **Aaron D. Robinson**.

One of us (C.K.G.) teaches leadership skills and works with troubled departments. At almost every session, someone will sidle up, curious about a case study: they want to know how what happened at their university came to be known externally. Of course, it didn't.

From what we've observed as a former university administrator and consultant (C.K.G.) and as a graduate student and working professional (A.D.R.), toxic research environments share a handful of operational flaws and cognitive biases. Researchers and institutional leaders must learn how these infiltrate their teams, and tailor solutions to keep them in check.

People who enter research generally share several values. Honesty, openness and accountability come up again and again when C.K.G. asks researchers to list what makes a good scientist. The US National Academies of Sciences, Engineering, and Medicine says that these values give rise to responsibilities that "make the system cohere and make scientific knowledge reliable"[1]. Yet every aspect of science, from the framing of a research question through to publication of the manuscript, is susceptible to influences that can counter good intentions.

C.K.G. coined the mnemonic TRAGEDIES (Temptation, Rationalization, Ambition, Group and authority pressure, Entitlement, Deception, Incrementalism, Embarrassment and Stupid systems) to capture the interlocking factors that can lead scientists astray[2] (see 'A table of tragedies').

LAB HEALTH
A *Nature* special issue
nature.com/collections/labhealth

Consider this true story. A professor asked a beginning graduate student to verify that numbers on a data sheet matched those in a figure in a scientific manuscript, and to state in an e-mail that the data were accurate as far as he could tell. The paper described work that had been completed before the student arrived on campus and about which he knew little. Later, the student discovered that the paper was submitted the day he sent his confirmation e-mail — and that he was listed as a co-author. We can imagine his reactions.

He might be tempted to let the inappropriate authorship stand to gain a publication and avoid confronting his adviser. He could rationalize that he was new and the

▶

professor knew what was appropriate. He could feel ambition to get ahead, and pressure from an authority figure, and he would be aware of some deception — he knew he didn't qualify for authorship because he had not been involved in the research in any substantive way. There's no evidence that he disputed his inclusion as an author.

He was almost certainly unaware that his name had been added because reviewers who rejected a previously submitted version of the manuscript had questioned whether a single researcher could have done the work. Later, an investigation found that the professor had orchestrated the paper's publication with "falsifying intent" to suggest that a different publication had been independently confirmed. We know about his dilemma because of a misconduct investigation that underwent congressional scrutiny. As far as we can tell, that student is no longer in science.

Here's another real example. A research coordinator decided to resign in the face of problematic data. Before leaving, she told a postdoc that some scans had been done on only 6 of the 50 subjects in the data set — and that the results did not support the hypotheses of the principal investigator[3]. The tragedy here is that the postdoc didn't have the resources to work the dilemma through and instead simply looked for a new job. Again, pressure from an authority figure and potential embarrassment made it challenging to take appropriate action.

Here's a situation that wraps everything together. An assistant professor knows that her paper is more likely to have higher citation counts, which her institution values, if she includes as an author a senior person in her field who didn't contribute to the paper other than a passing lunch discussion. The tragedies here encompass temptation, rationalization, ambition, authority pressure, deception, stupid systems — and maybe entitlement if she's working hard and feels that the ends justify the means. If she does it once, and gets rewarded, how likely is she to do it again? (This is incrementalism.)

A TABLE OF TRAGEDIES

Nine common pitfalls

The factors that lead to bad decisions can be represented by the mnemonic TRAGEDIES. Here are some examples of each pitfall. Recognizing these and responding appropriately can save a career and strengthen science.

Temptation
"Getting my name on this article would look really good on my CV."

Rationalization
"It's only a few data points, and those runs were flawed anyway."

Ambition
"The better the story we can tell, the better a journal we can go for."

Group and authority pressure
"The PI's instructions don't exactly match the protocol approved by the ethics review board, but she is the senior researcher."

Entitlement
"I've worked so hard on this, and I know this works, and I need to get this publication."

Deception
"I'm sure it would have turned out this way (if I had done it)."

Incrementalism
"It's only a single data point I'm excluding, and just this once."

Embarrassment
"I don't want to look foolish for not knowing how to do this."

Stupid systems
"It counts more if we divide this manuscript into three submissions instead of just one."

TAKE RESPONSIBILITY

Departments and institutions might protest that there is little they can actually do: the funding and recognition system itself favours poor methods and can lead to "the natural selection of bad science"[4]. We respond that institutional leaders — from those who supervise students to presidents and chancellors — must take responsibility for the working environment at their organizations.

There are two fundamental steps to improve the situation that are completely under local control. One is assessing empirically the integrity of research cultures. The second is developing research ethics education that is relevant to and integrated with how trainees actually learn to do science.

Unfortunately, most education provided on the responsible conduct of research, at least in the United States and Canada, focuses almost exclusively on compliance. Few students need to be able to define fabrication formally or to identify relevant sections of the Belmont Report, the 1978 document codifying how to treat human subjects of research.

What they really need is information about how to take action and to make decisions in tricky circumstances. And how to approach a senior faculty member or colleague over concerns about data in a constructive, non-threatening manner. And how to identify people who can give useful, disinterested advice. And how to blow the whistle.

Researchers such as Michael Mumford at the University of Oklahoma in Norman have found that effective programmes give students multiple strategies for analysing situations to identify ethical issues. They encourage interaction and provide real case studies of positive and negative examples with emotional impact — not just regulations or guidelines[5]. Students learn more than rules; they rehearse strategies for responding to tough cases and for anticipating consequences.

Context is as important as content. Courses on the responsible conduct of research are often outsourced or run online, which underscores the low priority of this instruction. Instead, courses should be taught by scientists within trainees' disciplines and run for more than a single session. For example, a programme at the University of Kansas in Lawrence for engineering students has 15 hour-long sessions with guest lectures by active faculty members. The University of California, Berkeley, incorporates topics on responsible conduct alongside experimental design and statistics. Department heads and lab leaders should also integrate a wider range of issues around research integrity — including mentoring, methods courses and career seminars —into group meetings, seminars, journal clubs and any event at which researchers discuss how science is done.

Many faculty members will feel ill-equipped for these discussions, or worry that it will take up precious time. But if researchers and institutional leaders want to support the most rigorous research, or even just forestall scandals, they must make this commitment. A small first step is to acknowledge that TRAGEDIES exist and discuss strategies to check them. A simple way to approach the topic is by talking regularly about their own dilemmas and how they resolved them, successfully or not. Any faculty member who conducts research, submits proposals, reviews manuscripts or works as an editor will have anecdotes that trainees can learn from.

Beyond that, there are a wealth of case studies and instructional materials that have been compiled through EthicsCORE, the National Academy of Engineering Online Ethics Center and the Committee on Publication Ethics. Several societies also offer relevant materials online, including the American Geophysical Union, the American Physical Society, the American Society for Cell Biology and the Society for Neuroscience.

ASSESS THE CLIMATE

Even exemplary training will not alter a toxic work environment. The informal curriculum — what researchers observe about how work is actually done — will always trump a formal one. To support research integrity, institutions must get a handle on what their local informal curriculum is teaching, and that means evaluating the current research environment.

There are many ways to gather data to make change. A quick self-assessment using our Academic Unit Diagnostic Tool (AUDiT) (go.nature.com/2jliagk), can help to surface indicators of vibrancy and problems in a unit's culture. Conducting 'exit interviews' with

ILLUSTRATION BY DAVID PARKINS

PhD students, postdocs and professors, and looking for patterns, has also proven valuable, as have institution-wide or department-wide surveys about student and staff experiences.

The only validated tool we know of in this area is the Survey of Organizational Research Climate (SOURCE). It assesses seven dimensions, including integrity norms, adviser–advisee relations and departmental expectations. Results correlate with self-reported rates of detrimental research practices: institutions with low scores of integrity norms will also tend to have higher levels of reported fraud and sloppy record keeping[6].

The survey can be done online in 15 minutes, and responses are aggregated to ensure individual confidentiality but still show differences across groups. That can help to identify both pockets of good practice and areas needing improvement. One large institution in the midwestern United States has used results to prompt faculty members within specific departments to talk more with graduate students about authorship, peer review and data management.

As well as being used to compare departments across an institution, the results can be compared against anonymous benchmarking data aggregated by the National Center for Professional and Research Ethics at the University of Illinois at Urbana-Champaign (which C.K.G. runs). Now no one can retort, "well, all departments in our field are that way".

The management literature is clear that one powerful way to bring systemic organizational change is to find 'bright spots' — systems or places in an organization that are working well — study them and seek to spread their successful practices. For that, we need data on where the bright spots are, and the will to act.

The solutions are straightforward, if not necessarily simple. ■

C. K. Gunsalus *is the director and* **Aaron D. Robinson** *a collaborating expert on design and research at the National Center for Professional and Research Ethics at the University of Illinois Urbana-Champaign.*
e-mail: gunsalus@illinois.edu

1. National Academies of Sciences, Engineering, and Medicine. *Fostering Integrity in Research* (National Academies Press, 2017); available at https://doi.org/10.17226/21896
2. Gunsalus, C. K. *The Young Professionals' Survival Guide* (Harvard Univ. Press, 2012).
3. Sieber, J. E. *J. Empir. Res. Hum. Res. Ethics* **7**, 3–14 (2012).
4. Smaldino, P. E. & McElreath, R. *R. Soc. Open Sci.* **3**, 160384 (2016).
5. Mumford, M. D. *et al. Ethics Behav.* **18**, 315–339 (2008).
6. Martinson, B. C., Thrush, C. R. & Crain, A. L. *Sci. Eng. Ethics* **19**, 813–834 (2013).

Go beyond bias training

Ambiguity in expectations and evaluations harms progress, say **Rodolfo Mendoza-Denton** and colleagues.

One morning in February 1934, the police showed up at J. Robert Oppenheimer's home in Berkeley, California, to ask why he had left his date in a car by herself all night. Oppenheimer explained that he had gone for a stroll, got lost in his thoughts and walked home, forgetting his car and companion.

Newspapers reporting this story for Valentine's Day revelled in tales of the absent-minded professor, an archetype that most of us recognize. Brilliant, but short on social graces, such thinkers are assumed to be too busy pondering the deepest questions of the Universe to be ▶

bothered with the quotidian.

This archetype, however, can also give licence to the neglect of students. Professors are often excused from knowing the requirements and timelines of graduate programmes. Graduate students regularly receive minimal guidance. The underlying supposition is that the path to success will reveal itself if the student 'has what it takes'. Lack of direction is often deemed a litmus test for the brilliance of the student.

In our view, women and under-represented minorities face a double whammy under these conditions. First, ambiguous expectations and guidelines allow bias to influence professors' judgements of student work. Second, environments with unclear or inchoate norms can depress the performance and progress of students in marginalized groups, further perpetuating notions of who qualifies as 'brilliant'.

Interventions designed to address disparities in science focus largely on changing individual attitudes[1]. Our surveys of science, technology, engineering and medicine (STEM) departments at the University of California, Berkeley, suggest another, complementary target: the structure of the training programmes themselves, and the cultures built around them.

AMBIGUITY AND BIAS

Professors are generally in control of deciding which of their students' research is nurtured, funded and eventually published. And, like all individuals, professors' judgements are subject to bias.

Ambiguous cues about trainees' and candidates' performance allow evaluators to incorporate their own, often unconscious, expectations into their assessments. Presented with job applications designed to represent credible but not stellar candidates (would-be research assistants with a published paper and two years of work experience, but low academic achievement), faculty members rated the same work and credentials more positively when it was accompanied by a male name than when associated with a female name[2]. A similar study found that research abstracts were rated as being of higher quality if presented as being authored by men and on topics, such as computer-mediated communication, that tend to be associated with males[3]. By contrast, another study showed that when candidates for a faculty job were presented as equally strong, according to numerical ratings presumably made by other faculty members, traditional gender biases were reversed, at least in this instance[4]. Similar patterns have been seen with respect to race[5]. In one study[5], evaluators showed strong preferences towards white candidates when the candidates' qualifications were ambiguous, but no preference when candidates were unambiguously strong or weak. In short, ambiguity serves as fertile ground for the expression of bias.

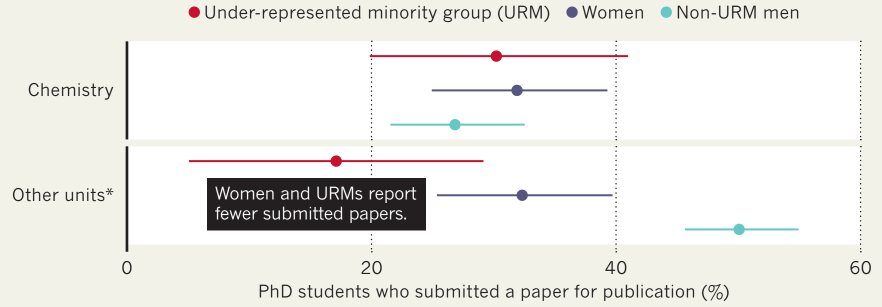
> **"Structured programmes need not be impersonal or automated."**

In addition, it dents the performance of those under evaluation. In one of our studies, about 150 women were asked to wait in one of three rooms arranged to reflect attitudes of their purported evaluator[6]. In one setting, to suggest that the evaluator held sexist views, the decor included a poster of a bikini-clad woman. In a second, the decor featured a volunteering award with a logo promoting equality to suggest that the evaluator advocated gender equality. The third room was ambiguous, with a banner from a university and a certificate for volunteering in the 'Ivy League Undergraduate Division'. (A separate survey confirmed the rooms gave the desired impressions.)

We assessed study participants' concerns about gender-based discrimination with an openly available, previously developed instrument[7]. Concerns did not affect test performance in either the chauvinist or progressive conditions: these groups answered about 8 of 12 moderately difficult analogies correctly. But in the ambiguous room, women who were concerned about being the target of prejudice averaged fewer than 7 correct answers, a strong effect.

NO GAP IN CHEMISTRY

The College of Chemistry stands out from some other units at the University of California, Berkeley, because it shows no difference in submission rates across various groups.



Women and URMs report fewer submitted papers.

*Division of Mathematical and Physical Sciences and Department of Electrical Engineering and Computer Science.

SOURCE: REF. 8

BRIGHT SPOT

A key requirement for advancement in academia is publication. Almost every step of that process — which project to encourage, how to allocate resources and credit, where and when to submit a manuscript — involves uncertainty. In a survey that R.M.D. conducted with Berkeley psychologist Aaron Fisher and his colleagues across the university's Division of Mathematical and Physical Sciences, the Department of Electrical Engineering and Computer Science and the College of Chemistry[8], graduate students were asked whether they had been an author on a paper submitted for publication in the past year. This reflects scholars' participation in the research enterprise independent of the vagaries of manuscript review and acceptance.

Response rates for men who were not in under-represented minorities and for all students in the engineering and mathematics departments were around 40%. Those for women, people in under-represented groups and the chemistry college were just over 50%. Women and people in under-represented minority groups had fewer submissions than did their white or Asian male counterparts, even when controlling for factors such as time in the programme, advancement to PhD candidacy and teaching responsibilities. To our surprise, however, race and gender did not predict the likelihood of publishing for people in the chemistry college (see 'No gap in chemistry').

Intrigued, we went on to examine 15 years' worth of data from Berkeley's PhD exit survey, which boasts a completion rate of 98%. The survey includes the questions: "Were you encouraged by faculty in your department to publish?" and "did you deliver any papers at national scholarly meetings?". The latter is often a precursor to publication.

Again, we found that, overall, women and under-represented minorities were much more likely than white and Asian men to answer 'no' to both questions across STEM fields. There were no statistically significant differences in the College of Chemistry.

What is happening in the college that seems to be levelling the playing field? Noting that the chemistry programme has been independently recognized for placing women with PhDs into elite faculty positions[9], our team has embarked on a series of interviews with faculty members, alumni and student advisers in this unit to identify some of the factors that could be fuelling the success of women and minority groups there. We are also interviewing people in the mathematics and physics departments, where we see strong disparities.

THREE HALLMARKS

Although preliminary, our data suggest that the chemistry college has the following characteristics.

Setting clear expectations from the start can help all trainees to thrive.

JIM WILSON/NYT/REDUX/EYEVINE

Advancement processes and procedures are clearly defined and systematically applied. For example, every student is expected to regularly present their research to colleagues and peers, including at a departmental seminar in the second year. This sets a public norm for productivity and affords multiple opportunities to learn from peers and near-peers as they themselves meet these expectations.

Student progress is overseen by multiple faculty members. For example, in each broad sub-field of the department, one adviser actively manages students' progress through the early stages of the programme, including helping to match students with research advisers. In addition to the academic and research advisers, each graduate student is entitled to a departmental 'associate adviser' once they pass their qualifying exam. This process ensures that students don't fall through the cracks, and engages multiple faculty members in collegial feedback as the student moves through the programme.

There is department-wide agreement about expectations for advancement. There are written guidelines for when students must choose an adviser, deliver seminars and pass qualifying exams. The expectation to publish is promoted officially. Before a student takes a qualifying exam, for example, advisers fill out a form that includes their assessment of when students will submit a paper, establishing that this is a norm, and prompting discussion.

These three observations suggest that requirements and regulations might not be enough. Rather, the community's knowledge, implementation and even application of standards are crucial to creating a culture in which students know what they need to do, and advisers know what they should encourage. We refer to this as a culture of structure.

RAPPORT AND STRUCTURE

Structured programmes need not be impersonal or automated. The twists and turns of discovery, and of people's lives, demand flexibility and exceptions. Our research suggests that trust between lab heads and lab members is essential — particularly in mentoring relationships that are interracial.

One of us (R.M.D.) and his colleagues published a study earlier this year intended to model interracial mentoring[10]. Some participants were asked to play the part of mentor, giving feedback on a speech for which a trainee had just three minutes to prepare. Trainees were asked to rate the quality of the feedback. In reality, all 'interactions' occurred through a video-chat in which one member of the pair was an actor performing a pre-recorded script.

Before the speech, half of the pairs were assigned an activity in which participants took turns asking and answering questions that escalate in self-disclosure — an exercise known to increase feelings of rapport. The control group took turns reading passages of novels to each other[11].

According to independent coders — who did not know which activity preceded the feedback session — trainees in pairs assigned to the rapport-building task gave better speeches and mentors provided warmer and more helpful feedback than did those in the control group. This held true for both same-race and interracial pairings.

Of course that's just one study, and more research is needed. But if the tenor of manipulated, short-term 'mentoring' can affect performance and feedback, it seems likely that the tenor of a trainee–adviser relationship could, too. In many STEM departments, emotions and feelings are deemed distractions. Our research suggests, instead, that establishing trust could be a key way to boost performance and parity through the ability to value each other.

In sum, our findings suggest fresh ways of interrupting bias in STEM education. Departments should adopt transparent policies and expectations for student progress that are communicated clearly to all. Professors and mentors should take time to build trust and rapport with students.

It is time we laid to rest the 'see you in five years' model, rooted in the specious notion that brilliance will find a way. Brilliance is most reliably nurtured through structure and trust. ■

Rodolfo Mendoza-Denton *is professor of psychology and executive associate dean of diversity and inclusion;* **Colette Patt** *is assistant dean in the mathematical- and physical-sciences division; and* **Mark Richards** *is professor of Earth and planetary sciences at the University of California, Berkeley.*
e-mail: rmd@berkeley.edu

1. Moss-Racusin, C. A. *et al. Science* **343**, 615–616 (2014).
2. Moss-Racusin, C. A., Dovidio, J. F., Brescoll, V. L., Graham, M. J. & Handelsman, J. *Proc. Natl Acad. Sci. USA* **109**, 16474–16479 (2012).
3. Knobloch-Westerwick, S., Glynn, C. J. & Huge, M. *Sci. Commun.* **35**, 603–625 (2013).
4. Williams, W. M. & Ceci, S. J. *Proc. Natl Acad. Sci. USA* **112**, 5360–5365 (2015).
5. Dovidio, J. F. & Gaertner, S. L. *Psychol. Sci.* **11**, 315–319 (2000).
6. Mendoza-Denton, R., Shaw-Taylor, L., Chen, S. & Chang, E. *J. Exp. Soc. Psychol.* **45**, 275–278 (2009).
7. London, B., Downey, G., Romero-Canyas, R., Rattan, A. & Tyson, D. *J. Pers. Soc. Psychol.* **102**, 961–979 (2012).
8. Mendoza-Denton, R. *et al. PLoS ONE* **12**, e0174296 (2017).
9. Laursen, S. L. & Weston, T. J. *J. Chem. Educ.* **91**, 1762–1776 (2014).
10. Leitner, J., Ayduk, Ö., Boykin, C. M. & Mendoza-Denton, R. *PLoS ONE* **13**, e0194123 (2018).
11. Page-Gould, E., Mendoza-Denton, R. & Tropp, L. R. *J. Pers. Soc. Psychol.* **95**, 1080–1094 (2008).

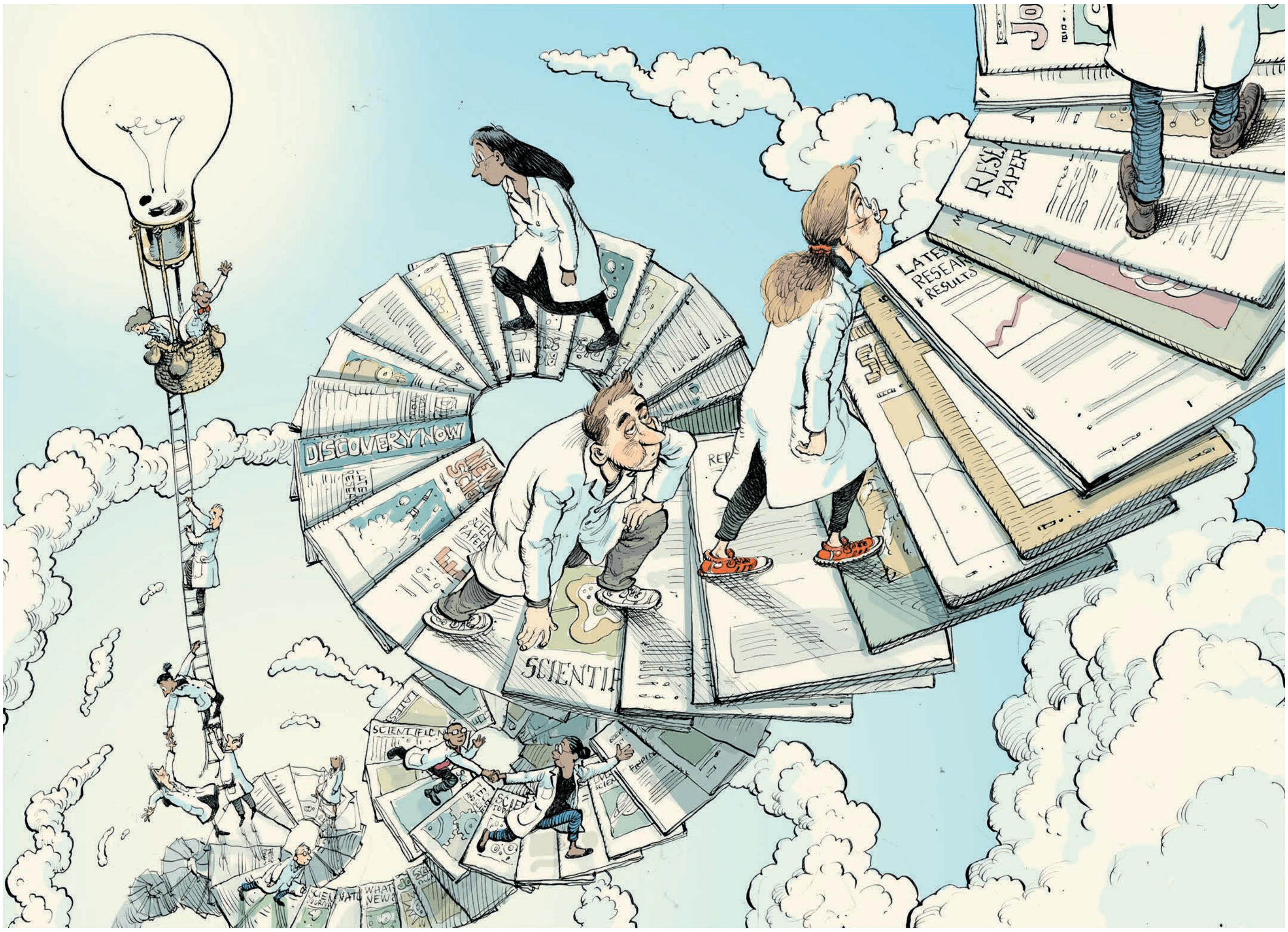


ILLUSTRATION BY DAVID PARKINS

Health tips for research groups

Nature asked scientists to recommend one thing that institutional and laboratory leaders could do to make science more productive, rigorous and happy.

DAVID NORRIS

Make lab health someone's job

Director of research, Donders Institute for Brain, Cognition and Behaviour, Radboud University, Nijmegen, the Netherlands.

Just over two years ago, our institute created a new position: sustainable-science officer. The job is to improve the working environment for everyone at our research centre. The effort succeeded.

The impetus grew from an internal colloquium about the future of science that took a surprising direction. More than 100 scientists attended, and both senior and junior researchers opened up about how trapped they felt in the current system of chasing publication credit to secure career prospects. Discussion spilled well over the time allotted. It was like people had finally been given a forum to say what they had previously kept to themselves.

The institute's directors decided to take action. I was then chair of the board, and we set out to identify what we could do ourselves, without waiting for action by funders or journals.

I canvassed the institute, e-mailing our 600 researchers an analysis of the discussion and asking for their thoughts. Problems fell into about 20 themes, including a disconnect between PhD graduation and academic positions, lack of incentives for multidisciplinary collaborations and a reward system that can penalize high-quality science. We formed workgroups around each theme to set tangible goals. Once people realized that the leadership was prepared to act, there was no shortage of volunteers.

Next, we crafted the role of science

sustainability officer, and hired Claudia Lüttke, a graduate student who had moved into science management. Her job, just over half-time and supported by internal funds, was to hold us accountable to our ideals. She was also someone whom trainees and faculty members could approach to talk through concerns and ideas.

The changes were broad (see go.nature.com/2kt3ka3). The institute, backed by the university, created a data-management programme to share and archive experimental results. We now assess scientific output on the basis of quality and do not have quantity as an explicit criterion. To encourage team science, we allow publications to feature in more than one student's PhD thesis, provided that joint authors made clear and distinct contributions. We established peer coaching teams of 4–6 researchers, run by volunteer facilitators who receive professional training. We instituted career-development plans for postdocs that give them the time and resources to devote to their own future. Everyone at the institute can access designated contacts outside their lab group to discuss good scientific practice informally (generally someone who is also an assistant or associate professor).

Even more importantly, we established a culture of openly discussing community issues across lab groups and among junior and senior researchers. Topics such as burnout, data-sharing dilemmas and PhD pressures have featured alongside neuroscience topics at weekly and monthly all-institute seminars. At meetings set up for trainees, principal investigators (PIs) talk about issues such as how they maintain work–life balance or make decisions about hiring postdocs. Our graduate students have been surprised that qualities such as enthusiasm and insight during job interviews could outrank the prominence and quantity of papers.

Unless specific people are charged by the institution with specific duties, sustainable science will not sustain itself. Lüttke has moved on to become a policy officer at another university, and her former role has been incorporated explicitly into a new senior position.

I believe that our scientific output is now as good or better than it was under the conventional system. It certainly generates less stress and wasted effort.

ULRICH DIRNAGL

Train the PIs

Professor of neuroscience, Charité University Medicine, Berlin.

I teach statistics, experimental design and good scientific practice to PhD students and postdocs. They are all fascinated with their scientific projects, and simply want to do 'the right thing' regarding the reproducibility, robustness and rigour.

Too often, my students tell me how steps to improve the validity of their work are obstructed by their supervisor or group leader. Quotes from their PIs include "I have published in *Science* and *Nature*"; "This will jeopardize our chances for acceptance"; or "This would take longer, and we might get scooped".

The importance of some techniques I teach, such as blinding, randomization and transparent analysis, are only now being emphasized in basic research. Methods for storing and analysing data have transformed in the past ten years. It makes no sense that senior researchers are exposed to this only haphazardly, if at all.

> ***"Let's start with mandatory courses in basic statistics and open science."***

Scientists have specialist training, but not a regulatory body or code of professional ethics. Nor do we have mandatory continuing professional development — as do lawyers, vets, nurses and even football referees. Pilots keep their licence only if they document a certain number of hours in flight and undergo flight review. Physicians must participate in continuing medical education. Should we consider similar requirements for academic scientists?

Let's start with mandatory courses in basic statistics and open science for group and departmental leaders in biomedicine. They could attend the same course I teach. This would help a lot, even if it only empowers the trainees.

MICHAEL J. ZIGMOND

Show that you care

Professor emeritus, Department of Neurology, University of Pittsburgh, Pennsylvania.

Since the 1980s, I have co-taught a course on 'survival skills' for lab members. I wish I could have also taught a companion course for lab heads: 'basic care of lab members'.

Such a course would hang on a simple dictum: show the members of your lab — trainees and staff — that you care. Demonstrate in multiple ways that you consider yourself a mentor, and not just a supervisor.

That means providing clear guidelines on the relative roles that you and each trainee will have in selecting and executing research projects, on proper data management (including the maintenance of a lab notebook), on authorship, on support for attending meetings and on what it will take for you to feel that trainees are ready to move on.

I would also encourage you to start with a high level of supervision, then gradually allow more independence, checking back to make sure that all is going well and making it clear that you will be available whenever needed — even when travelling. Be sure to tell trainees to take an active role in their own training, including giving you feedback on how you might be more helpful.

Lab meetings are a time when you can establish your relationship with lab members, both as individuals and as a group. You can show that you value their ideas and consider your lab as an interactive community, not simply as a set of relationships between you and them.

Take time to celebrate successes — breakthroughs in the lab, presentations made, papers accepted. Praise in public, but keep strong criticism private and focused on specific actions or decisions, not on the individual.

And finally, the most important lesson I would teach? That the key to having a successful, productive lab — and to foster individuals who will go on to successful careers — is to make clear from the beginning that you are paying attention to lab members' welfare and their progress as scientists, not just to their specific projects and their contributions to your own ambitions.

KATHERINE THOMPSON-PEER

Get lab members multiple mentors

Physiology postdoctoral scholar, University of California, San Francisco.

Anyone designing academic research from scratch would not create a system in which PIs have so much control over their lab members' research decisions and career opportunities. It bakes in too much fragility, with too few checks and balances to detect poor decisions and avoid bad situations. We must encourage trainees to seek out extra mentors.

Academia has been moving away from the model that a single mentor will be a PhD student's guru. Thesis advisory committees have been implemented because it is in the best interest of the student and the science to have multiple eyes and perspectives. Postdocs need that, too.

In my experience as a young postdoc, I

knew that I should be seeking extra mentorship but never seemed to make time for it. It was only when I applied for a grant intended to help postdocs transition to independent positions that I began actually meeting with faculty members outside my own lab. Input from people besides my PI helped me to refine grant and job applications, and provided insight beyond either of our expertise.

Not everyone embraces the idea that it is essential for a postdoc to talk to faculty members other than their advisers about their science and career. Some advisers are possessive, and some trainees are worried about overstepping, imposing or causing offence. Advisers should set the tone. They should be checking in with their postdocs regularly, asking who they have met with, and offering introductions.

One mentor cannot meet all the needs of any trainee or junior colleague. New trainees should feel they are joining the university and not just the lab. They should be expected to build a network of mentors as well as collaborators — people with varied strengths willing to look at how they are collecting and analysing their data and designing their path forward. Department heads should ensure that everyone is informed

> *“One mentor cannot meet all the needs of any trainee or junior colleague.”*

TRACY T. CHOW

Be explicit about expectations

Postdoctoral fellow at the University of California, San Francisco.

Shortly after I joined the lab for my PhD, the joint heads called a special meeting. I was apprehensive. From everything I knew about how labs worked, formal meetings were usually requested by trainees, not investigators. It turned out that the content of that meeting had nothing to do with the particular scientific questions that drew me into the lab — how the ends of chromosomes are maintained. But it had a major, and positive, effect on my work as a scientist.

In retrospect, some of the topics seem trivial: we went into minute details on various ways to organize results of multiple experiments conducted simultaneously, and how to document where to find raw files. This conversation demonstrated how seriously my advisers took the need to track experimental details and progress, and that they expected every experiment to clearly state the purpose, experimental conditions, results and conclusions, as well as plans for the next experiments.

Something else happened during that meeting, too. Trainees were encouraged to talk about their expectations for the mentors, and to share what was and wasn’t working. This also built a stronger bond between lab mates, and made us more comfortable approaching each other for future conversations.

As researchers, our conversations with advisers dwell so much on specific scientific ideas, specialized techniques and data interpretation that we sometimes omit discussions about what makes a great paper in our field, or how to plan a project that can both meet specific goals and be open to serendipity. Without realizing it, we end up making our best guesses on how to do good science rather than making time to discuss it. I am grateful that one such discussion occurred as early in my career as it did. That formal meeting to set expectations increased my self-awareness of how I work as a scientist and boosted my scientific insight. It taught me how easy it is to assume that others think similarly to us — which often is not the case — and reminds me how important it is to make assumptions explicit, for others and for ourselves.

Over time, I have found ways to ensure that these essential conversations happen. For example, my postdoctoral fellowship required me to have a statement of scientific and professional growth signed off by my adviser each year. I used this opportunity to share my expectations and a trajectory for my career development. Whenever I communicate with my adviser (in person, or through Skype or e-mail) about project updates, new data or experimental plans, I work to keep the conversation going two ways, and indicate what feedback I’m seeking. For instance, I list ideas for next steps, state what I think is most promising and ask what they think is feasible. In lab and committee meetings, I ask for input about expectations and timelines for the work to be presented at a conference or submitted for publication, and whether it would be appropriate to reach out to a possible collaborator. I also consider when my advisers are likely to be most approachable. For one, I might want to start a conversation during a short walk to a seminar. Some are more open to discussion over morning coffee; others in the afternoon.

That invaluable conversation I had so long ago came about through happenstance: two lab heads realized that their disparate working styles might cause confusion. But it taught me how important it is that advisers and trainees make sure they happen. Conversations about lab expectations should occur as a matter of course, not chance. ■

A ward in the Am Spiegelgrund clinic in Vienna, in the 1940s.

COURTESY OF THE DOCUMENTATION CENTRE OF AUSTRIAN RESISTANCE.

AUTISM

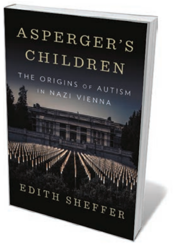
Hans Asperger's Nazi collusion

Simon Baron-Cohen absorbs the grave revelations in a study on a paediatrician enmeshed in autism's history.

The Austrian paediatrician Hans Asperger has long been recognized as a pioneer in the study of autism. He was even seen as a hero, saving children with the condition from the Nazi killing programme by emphasizing their intelligence. However, it is now indisputable that Asperger collaborated in the murder of children with disabilities under the Third Reich.

Historian Herwig Czech fully documented this in the April 2018 issue of *Molecular Autism* (a journal I co-edit; see H. Czech *Mol. Autism* **9**, 29; 2018). Now, historian Edith Sheffer's remarkable book *Asperger's Children* builds on Czech's study with her own original scholarship. She makes a compelling case that the foundational ideas of autism emerged in a society that strove for the opposite of neurodiversity.

These findings cast a shadow on the history of autism, already a long struggle towards accurate diagnosis, societal acceptance and support. The revelations are also causing debate among autistic people, their families, researchers and clinicians over whether the diagnostic label of Asperger's syndrome should be abandoned.



Asperger's Children: The Origins of Autism in Nazi Vienna
EDITH SHEFFER
W. W. Norton (2018)

In 1981, psychiatrist Lorna Wing published the paper in *Psychological Medicine* that first brought Asperger's clinical observations to the attention of the English-speaking medical world, and coined the term Asperger's syndrome (L. Wing *Psychol. Med.* **11**, 115–129; 1981). A decade later, in the book *Autism and Asperger Syndrome* (1991), developmental psychologist Uta Frith translated into English the 1944 treatise by Asperger in which he claimed to have discovered autism.

Finally, in 1994, the American Psychiatric Association (APA) recognized the diagnosis of Asperger's syndrome in the fourth edition of its *Diagnostic and Statistical Manual* (*DSM*). The syndrome is characterized by strengths such as unusually deep, narrow interests, and challenges in social communication and interaction, in people with average IQ or above and no history of language delay. (In the 2013 revision of the *DSM*, the APA deleted Asperger's syndrome in favour of a single category, autism spectrum disorder.)

In digging anew into the deeper historical context of Asperger's work, Sheffer fills in parts of the story anticipated in John Donvan and Caren Zucker's history of autism, *In a Different Key* (2016; see B. Kiser *Nature* **530**, 159; 2016), which referred to Czech's early findings. Sheffer reveals how the Nazi aim of engineering a society they deemed 'pure', by killing people they saw as unworthy of life, led directly to the Holocaust.

With insight and careful historical research, Sheffer uncovers how, under Hitler's regime, psychiatry — previously based on compassion and empathy — became part of an effort to classify the population of Germany, Austria and beyond as 'genetically' fit or unfit. In the context of the 'euthanasia' killing programmes, psychiatrists and other physicians had to determine who would live and who would be murdered. It is in this context that diagnostic labels such as 'autistic psychopathy' (coined by Asperger) were created.

Sheffer lays out the evidence, from sources such as medical records and referral letters, showing that Asperger was complicit in this Nazi killing machine. He protected children he deemed intelligent. But he also referred several children to Vienna's Am Spiegelgrund clinic, which he undoubtedly knew was a centre of 'child euthanasia', part of what was later called Aktion T4.

This was where the children whom Nazi practitioners labelled 'genetically inferior' were murdered, because they were seen as incapable of social conformity, or had physical or psychological conditions judged undesirable. Some were starved, others given lethal injections. Their deaths were recorded as due to factors such as pneumonia.

Sheffer argues that Asperger supported the Nazi goal of eliminating children who could not fit in with the *Volk*: the fascist ideal of a homogeneous Aryan people.

Both Czech and Sheffer include details on two unrelated children, Herta Schreiber and Elisabeth Schreiber, and their referral letters, signed by Asperger. In these, the paediatrician justifies Herta's referral to Am Spiegelgrund because she "must be an unbearable burden to the mother"; and ▶

Elisabeth's, because "in the family, the child is without a doubt a hardly bearable burden". These provide proof that he effectively signed their death warrants.

Nearly 800 children were killed in Am Spiegelgrund. Asperger went on to enjoy a long academic career, dying in 1980.

Both *Asperger's Children* and Czech's paper converge on the same conclusion. Personally, I no longer feel comfortable with naming the diagnosis after Hans Asperger. In any case, this is a category rendered moot in the most recent edition of the *DSM* (used in the United States). European nations will follow this diagnostic lead in 2019, with the 11th edition of the *International Classification of Diseases*.

The future use of the term, of course, is a discussion that must incorporate the views of autistic people. Many take pride in the term Asperger's syndrome as part of their identity, feeling it refers to their personality and cognitive style, which obviously do not change simply because of historical revelations. They might not, therefore, want a change. Others have already written about switching to using 'autism' (or autism spectrum disorder, or autism spectrum condition) to describe their diagnosis.

For brevity and neutrality, I favour the single term autism. However, because of the considerable heterogeneity among autistic people, I think it could be helpful for them and their families — together with autism researchers, clinicians and relevant professionals — to discuss whether subtypes should be introduced.

> ***"The future use of the term Asperger's syndrome is a discussion that must incorporate the views of autistic people."***

When Wing coined the term Asperger's syndrome, none of us was aware of Hans Asperger's active support of the Nazi programme. As a result of the historical research by Sheffer and Czech, we now need to revise our views, and probably also our language. *Asperger's Children* should be read by any student of psychology, psychiatry or medicine, so that we learn from history and do not repeat its terrifying mistakes. The revelations in this book are a chilling reminder that the highest priority in both clinical research and practice must be compassion. ■

Simon Baron-Cohen *is director of the Autism Research Centre at the University of Cambridge, UK, and president of the International Society for Autism Research.*
e-mail: sb205@cam.ac.uk

BIOTECHNOLOGY

Blood, sweat, tears and biotech

Eric Topol extols a gripping account of the rise and fall of US medical-testing company Theranos.

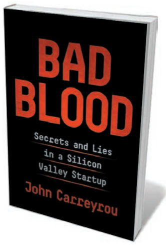
Few scandals have so gripped both the health-care and technology industries as the seismic rise and fall of blood-testing company Theranos. In *Bad Blood*, acclaimed investigative journalist John Carreyrou, who broke the story in 2015, presents comprehensive evidence of the fraud perpetrated by Theranos chief executive Elizabeth Holmes. Specifically, Holmes and the company's former president Ramesh 'Sunny' Balwani raised more than US$700 million through "elaborate, years-long fraud in which they exaggerated or made false statements about the company's technology, business, and financial performance", as the US Securities and Exchange Commission put it in March this year.

By the time Carreyrou's *Wall Street Journal* story and a long chain of follow-ups had led to regulators closing down Theranos's labs in 2016, nearly 1 million lab tests had been run in California and Arizona. A significant proportion of these were erroneous; all had to be voided. An untold number of people were harmed by the erroneous results: some underwent unnecessary procedures, received misdiagnoses of serious conditions and experienced emotional turmoil.

Carreyrou presents the scientific, human, legal and social sides of the story in full. Although some of it was previously reported in his extensive coverage, he unveils many dark secrets of Theranos that have not previously been laid bare.

The company's alluring goal, which changed as it evolved, was to quickly analyse a drop of blood for hundreds of different assays, at a fraction of prevailing costs. Collected in 'nanotainers' and allegedly tested in a diagnostic 'miniLab' the size of a microwave oven, the method was publicized as revolutionizing an industry that hadn't changed for decades. Holmes, who idolized Apple entrepreneur Steve Jobs, called it "the iPod of health care".

As a child in the 1990s in the United States, Holmes declared that she wanted to be a billionaire when she grew up. By age 30, she had achieved her goal. Highly intelligent, she had been accepted in 2002 to study chemical engineering at Stanford University in California as a President's Scholar, a prestigious programme that comes with a grant. She dropped out in her second year to start Real-Time Cures, the Palo Alto-based company that became Theranos and operated for 15 years.

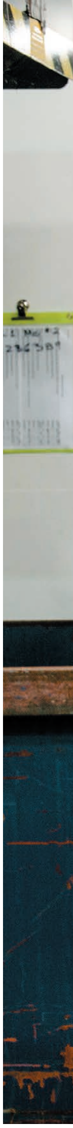


Bad Blood: Secrets and Lies in a Silicon Valley Startup
JOHN CARREYROU
Knopf (2018)

Carreyrou explores Holmes's talents and liabilities. Driven, and with an exceptional gift for selling ideas, she built up a board of high-level political figures, such as former secretaries of state George Shultz and Henry Kissinger, former secretary of defence William Perry, Marine Corps general (now secretary of defence) James 'Mad Dog' Mattis, and former senator Sam Nunn. Investors included media magnate Rupert Murdoch; partners numbered supermarket giant Safeway and pharmacy chain Walgreens. Channing Robertson, a professor of Holmes's at Stanford, was a board member and adviser to the company. In 2015, Holmes was recognized by then-president Barack Obama as a US ambassador for global entrepreneurship. The same year, vice-president Joe Biden sang her praises at a launch of the miniLab (which Carreyrou uncovers as completely fake; the lab was not operational at the time).

All the while, as Carreyrou reports, Holmes was lying about the nanotainer, contracts with the pharmaceutical industry and assay validation. She made false statements to the US Federal Drug Administration (FDA) and to the US government agency that regulates blood-testing labs, the Centers for Medicare and Medicaid Services. And she claimed that Theranos was being used on the battlefield in Afghanistan, saving soldiers' lives.

Holmes described the miniLab as "the most important thing humanity has ever built". But at best, the lab could do



Elizabeth Holmes, chief executive of Theranos, in a company facility in Newark, California.

CARLOS CHAVARRIA/NYT/REDUX/EYEVINE

immunoassays using microfluidics. The tiny blood sample had to be diluted extensively (for which there are no reference standards or precedents), leading to artefacts and spurious results. Later inspections by the FDA demonstrated poor quality control of multiple lab tests using Theranos equipment, and several examples of failed proficiency testing. The rest of the hundreds of routine assays the lab was supposed to deliver would require cytometry, general chemistry and DNA amplification. These were done using routine commercially available lab equipment, or were hived off to other facilities. That was the well-kept secret inside the toxic work environment that *Bad Blood* exposes.

Carreyrou describes firings and legal threats to and serious intimidation of employees, as well as industrial espionage involving spying on employees' social-media accounts. The book even goes into the suicide of a former leader of the company's chemistry group.

I met Holmes twice and conducted a video interview with her in 2013, for the medical-information website Medscape. At the time, I gave a fingerstick nanotainer blood sample and within 30 minutes received my results for many routine tests — allegedly showing, for instance, normal glucose and lipid levels in accordance with previous testing. Little did I know that they were run on a standard Siemens machine (I was not allowed to see the lab area) in the back room of Theranos, and had nothing to do with the miniLab. Like so many others, I had confirmation bias, wanting this young, ambitious woman with a great idea to succeed. The following year, in an interview with *The New Yorker*, I expressed my deep concern about the lack of any Theranos transparency or peer-reviewed research.

> *"An untold number of people were harmed by the erroneous results."*

Near the end of *Bad Blood*, Carreyrou describes how, in 2015, litigator David Boies — then Theranos's legal counsel — attempted to prevent *The Wall Street Journal* from publishing Carreyrou's reportage. For instance, Boies accused the paper of publishing Theranos trade secrets and making false and defamatory statements. Despite the $125 million invested in Theranos by Murdoch, the newspaper's owner, the pieces were published. We also learn about Carreyrou's tipster, a pathologist and blogger, along with so many employees who were rightfully afraid of hurting patients with fraudulent lab results. The combination of these brave whistle-blowers, and a tenacious journalist who interviewed 150 people (including 60 former employees) makes for a veritable page-turner.

My only criticism is the book's lack of reflection about lessons learnt from this debacle. How did a company rise to a valuation of $9 billion in a network of so many influential people, even as people were endangered? In my view, letting this technology loose (despite grand claims) without a single publication by independent scientists, never mind replication, was a recipe for jeopardy. Had the medical community and regulators held the company accountable, this could have been pre-empted. There have been other examples of Silicon Valley companies that rose meteorically, but none has put patients' health at risk. Hopefully, the evidence in *Bad Blood* will stop it happening again. ■

Eric Topol *is executive vice-president of the Scripps Research Institute in La Jolla, California.*
e-mail: etopol@scripps.edu

The 2013 United Nations Climate Change Conference in Warsaw.

KACPER PEMPEL/REUTERS

COMMUNITY

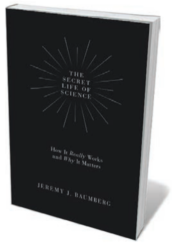
The scientific ecosystem

Jonathan Adams is invigorated by a study on the nitty-gritty of research, from publishing to conferences.

Experienced scientists have two core tasks. One is to produce great research. The other is to encourage people to think about how to do great research. Jeremy Baumberg has done the first; now he tries his hand at the second. He was a professor at 30 and is a fellow of the Royal Society. He is well-funded and experienced internationally, in both the private sector (at IBM and Hitachi) and the public realm (the University of Cambridge, UK), running nanotechnology groups with a focus on nanophotonics. As *The Secret Life of Science* shows, he's in an excellent position to create a refreshing description of how research "really works and why it matters".

In this solid, coherent and thoughtful study, Baumberg deploys an ecosystem model to describe research processes. (I was once an ecologist and know the usefulness of such models in capturing resource flows, competition, behavioural parallels and so on.) He clarifies the dynamics of publishing, conferences, media, translation to industry and careers, and provides a broad summary of influence.

The chapter on publishing will strike a chord for many *Nature* readers. He touches on how scientists must "clamor for attention", driving them to publish often and in high-impact journals. He notes information from, and influence of, publication metrics. But he doesn't discuss concerns about their misuse, as captured in the Declaration on Research Assessment launched in San Francisco, California, in 2012 and endorsed this year by the UK research councils. Nor does he mention the Leiden Manifesto on how to do things better (D. Hicks *et al. Nature* **520**, 429–431; 2015).



The Secret Life of Science: How It Really Works and Why It Matters
JEREMY J. BAUMBERG
Princeton University Press (2018)

Baumberg's take on conferences might be less familiar. He asserts that the conference system dominates science. Does it? Although I agree with him that topical conferences are the most valuable for interaction, I'm dubious about their importance in the life and social sciences, at least for leading researchers. Perhaps we need a new impact indicator for that part of the ecosystem?

The book offers intriguing insights into the grant system (the tensions around how to set goals and distribute funding for science) and technology-transfer problems (innovation in established enterprises). But Baumberg's vignettes — such as one on how established technology suppressed innovation at photography company Kodak — whetted my appetite for more detailed examples. Elsewhere, he draws attention to the policy challenge of training too many young researchers with too little funding per head, and comments on the over-supply of postdocs, in an echo of UK and US reports. How this can be managed, and how the situation is different in China — still in a growth phase — deserved much more space.

Baumberg's recommendations for change are admirably succinct. Rightly, in my view, he tackles the overabundance of scientists and the benefits of constraining the pipeline. He proposes changes in funding distribution to restore diversity of ideas. He calls for better conferences attended by fewer people, and for improving tools through artificial intelligence and customized knowledge retrieval. He commends open access and open data, and proposes open instrumentation.

I was pleased to see that he enjoins leaders to nurture more-considered postdoctoral careers and to ensure that organizations manage people and projects more consistently. The latter point came up repeatedly in an as-yet unpublished survey for the advocacy group Universities UK on international collaboration. A key benefit of the European framework research-funding programmes has been the establishment of sound, consistent regional bureaucracy.

I enjoyed much of *The Secret Life of Science*, but it has idiosyncrasies. It is definitely not about the whole research base. Baumberg suggests bringing in social scientists to help identify scientific ecosystem services (intuition, for example), yet he fails to point out that social scientists such as Robert K. Merton have influenced the structure and practice of science in the United States.

I wondered who this book is for. It lacks the accessible style and high-quality graphs and images that would be useful to the aspiring scientist. And for the academic, it has not a single reference, so it is impossible to follow sources and difficult to distinguish between Baumberg's original thinking and what he draws from reading. I found this problematic in sections touching on my research areas of tracking collaboration and impact through publishing. Some statements, such as those on citation patterns, are highly simplified, and there is no link to the substantial research background or reports such as *The Metric Tide* (2015). Elsewhere, the very extensive work on 'the science of science' in many countries, particularly that from the US National Science Foundation, is not recognized.

Baumberg sees *The Secret Life of Science* as providing a bird's-eye view that "scientists and public alike have lacked". I found his perspective more particular than that. But that does not detract from the validity of the narrative, nor the high value of the final discussion. ■

Jonathan Adams *is director of the Institute for Scientific Information, part of Clarivate Analytics, and a visiting professor at the Policy Institute at King's College London.*
e-mail: jonathan.adams@clarivate.com

Correspondence

Don't grant robots legal personhood

In an open letter, 156 artificial-intelligence experts from 14 European countries (go.nature.com/2t5mgov) have rejected the European Parliament's recommendation that robots should have legal status as electronic persons. This would make robots responsible for repairing any damage they might cause (go.nature.com/2wxlwg6). We are not signatories to the open letter, but endorse it nonetheless.

In our view, the parliament's recommendation is flawed. Its rationale seems to be that robots can be electronic juridical persons in the same way as companies are. But companies are constituted and run by real people. That is why they can be meaningfully attributed with intentions, plans, goals, legal rights and duties, and why they can be taught, praised or punished. Hence, they are considered to be responsible, accountable or liable for their actions.

Attributing electronic personhood to robots risks misplacing moral responsibility, causal accountability and legal liability regarding their mistakes and misuses. Robots could be blamed and punished instead of humans. And irresponsible people would dismiss the need for care in the engineering, marketing and use of robots. Even the Romans knew better: the owner of an enslaved person was fully responsible for any damage caused by that person (known as vicarious liability).

Luciano Floridi, Mariarosaria Taddeo *Oxford Internet Institute, University of Oxford, UK.*
luciano.floridi@oii.ox.ac.uk

Treating brainwaves is not an option

The rise in neurostimulation methods as potential treatments for neuropsychiatric diseases is encouraging (see *Nature* **555**, 20–22; 2018). As president of the International Pharmaco Electroencephalogram Society, however, I question your suggestion to investigate how brainwaves might be manipulated to benefit patients.

Brainwaves are just the by-products of ongoing neuroelectrical interactions, so an underlying condition gives rise to anomalous patterns on electroencephalograms. Clinical interventions that target the faulty cellular mechanisms responsible will alter those brainwave patterns. This holds true for potential neurostimulation therapies, such as flashing lights and pink noise that you discuss.

In my view, manipulating the brainwaves themselves risks confusing cause and effect, so is unlikely to work as a treatment.

Sebastian Olbrich *University of Zurich, Switzerland.*
sebastian.olbrich@puk.zh.ch

Local solar solutions can save water

Solar radiation management is good for more than just global geoengineering (see also A. A. Rahman *et al. Nature* **556**, 22–24; 2018). At the local level, the brightening of urban surfaces has helped to cool cities since the sixth century BC in Athens, and the same physics that causes aerosols to brighten the stratosphere (Mie theory) can also make water more reflective.

Dissolving air in water to generate bright, part-per-million dispersions of reflective microbubbles can curb evaporation by reducing solar warming (R. Seitz *Clim. Change* **105**, 365–381; 2011).

Aside from cooling urban reservoirs, this water-conservation technique allows farmers who cannot afford reservoir covers to reduce crop losses from drought.

I suggest that some of the money spent on arguing about global governance of geoengineering could be better used to develop local water-conservation technologies.

Russell Seitz *The Climate Institute, Washington DC, USA.*
russellseitz@gmail.com

Count the costs of economic skirmishes

China could decide to retaliate against proposed US tariffs on imported Chinese products (see go.nature.com/2wr6etx) by buying its goods from elsewhere. Such a move could have effects on environmental sustainability that need to be taken into account.

For instance, China might choose to import soya beans from Brazil. Brazilian farmers would then expand their production, putting further sensitive rainforest regions at risk. Yields would be maximized by overuse of fertilizers and pesticides, leading to regional contamination of soil and water.

Or there could be environmental benefits to switching. China might turn to Europe for its imports of civilian aircraft, for example, worth US$14 billion to the US market. Compared with the United States, European countries have higher environmental standards: they strongly support the Paris climate accord, and manufacturers there aim to reduce hazardous wastes and have production regulations that penalize environmental damage.

Germany, Japan and South Korea also have superior environmental norms and could step in as alternative suppliers of cars to China. Germany, for example, is heavily investing in renewable-energy resources.

In our view, proper management of the socio-economic and environmental outcomes is a crucial component of fair and stable trade policies (see also Y. Geng *et al. Ecol. Econ.* **132**, 245–254; 2017).

Yong Geng *Shanghai Jiao Tong University, Shanghai, China.*
Joseph Sarkis *Worcester Polytechnic Institute, Massachusetts, USA.*
ygeng@sjtu.edu.cn

Get facts straight on computer women

The photograph you captioned as showing female programmers of the Electronic Numerical Integrator and Computer (ENIAC) in the 1940s in fact dates from 1962, and the women were not ENIAC programmers (see W. R. Poster *Nature* **555**, 577–580; 2018; now corrected). They are Patsy Boyce (later Simmers), Gail Beck (later Taylor), Millicent Beck and Norma Stec, who worked at the US Army Ballistics Research Laboratory in Aberdeen, Maryland. The photo shows the progressive miniaturization of circuitry through four generations of computers: ENIAC (1945), EDVAC (1949), ORDVAC (1952) and BRLESC (1962).

Simmers and Stec were mathematicians. They used the ORDVAC and BRLESC computers to calculate US Army weapons-firing tables (see go.nature.com/2ruixtc). As far as I can tell, the stories of the other two women in the photo have not been told. Little is known about the computer artefacts they are holding.

The photo is also an example of historical casual sexism in the field of computing. When originally published (*Army Research and Development News Magazine* **3**, 4; 1962), it was captioned "BRLESC girls pictured below", playing on the pronunciation of BRLESC as 'burlesque'.

Ken Shirriff *Redwood City, California, USA.*
shirriff@righto.com

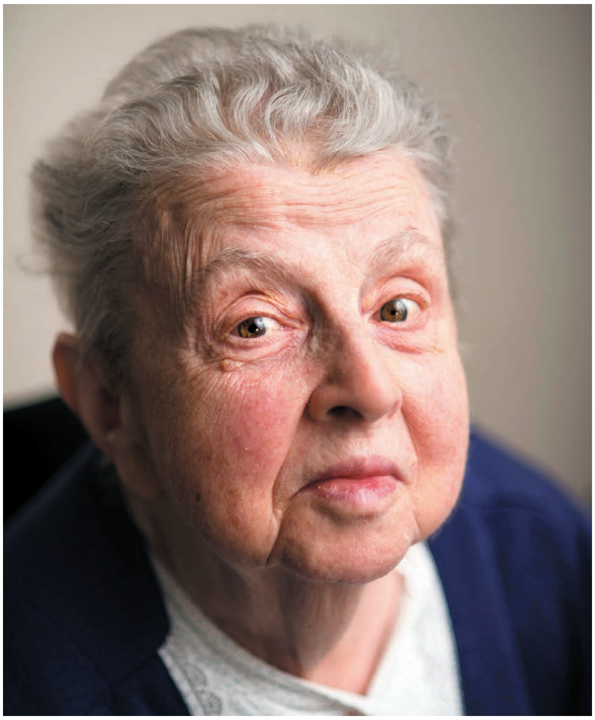
CORRECTION

The Nature Index article 'Facing down disaster' (*Nature* **555**, S66; 2018) misstated the length of the NIED ocean-floor fibre-optic network. It covers 5,500 kilometres, not 5,700.

Ruth S. Nussenzweig

(1928–2018)

Immunologist who paved the way to a malaria vaccine.



NYU LANGONE HEALTH

Ruth Nussenzweig's pioneering work was central to the ongoing effort to develop vaccines against malaria. Each year, the disease afflicts more than 200 million people and causes around 450,000 deaths, primarily in infants and young children in sub-Saharan Africa, according to the World Health Organization (WHO). Nussenzweig discovered a way of immunizing mice against malaria, using a technique which others later confirmed could work in humans. She also had a crucial role in identifying a good target protein for vaccination. She died on 1 April at the age of 89.

Born Ruth Sonntag in 1928, in Vienna, her parents were physicians. Soon after the Nazi occupation of Austria in 1938, her family fled to Brussels, where they hid for a year before moving to Brazil. Ruth received a degree in 1953 from the University of São Paulo School of Medicine, to which she returned to defend her PhD in 1968. At medical school, she met fellow student Victor Nussenzweig, and they married in the library. Their marriage of more than 60 years was also a lifelong scientific collaboration.

After a stint in Paris, Nussenzweig joined the New York University (NYU) School of Medicine in 1965, and turned her attention to immunizing mice with malaria sporozoites, weakened by radiation. Sporozoites are the infectious form of the parasite, introduced into humans by mosquito saliva; they migrate to the liver and reproduce, after which red blood cells become infected and symptoms begin. Nussenzweig's was an unconventional tack: she was attempting to actually prevent malaria infection from the outset. Most researchers were targeting the parasites only after they infected the blood cells.

Nussenzweig's approach proved fruitful. Within just two years, she published the seminal finding that mice could be protected from malaria by immunizing them with radiation-treated, or attenuated, sporozoites of *Plasmodium berghei*, the organism that causes the disease in rodents (R. S. Nussenzweig *et al. Nature* **216**, 160–162; 1967). In many of the immunized mice, the parasites were unable to reach the blood-infection stage. Before this, immunologists had deemed the strategy unlikely to be successful.

On the basis of this crucial observation, other researchers showed in the 1970s that human volunteers exposed to hundreds of bites from irradiated mosquitoes also developed immunity, and were protected from malaria infection (D. F. Clyde *et al. Am. J. Med. Sci.* **266**, 169–177; 1973). But it took several decades of work to isolate purified, radiation-attenuated sporozoites from mosquitos in a form that could be used as a vaccine in humans. This led to the development of the PfSPZ vaccine, which is undergoing phase II clinical trials in Africa to assess safety and efficacy.

Nussenzweig's next major breakthrough came in 1979. She showed, in collaboration with Victor, and their colleagues, that mice could also be protected from malaria infection using antibodies against one antigen — later named the circumsporozoite protein (CSP) (N. Yoshida *et al. Science* **207**, 71–73; 1980). This finding prompted Nussenzweig and others to identify and clone CSP from the parasites that infect humans: *Plasmodium falciparum* and *Plasmodium vivax*.

This research led other groups, along with the major UK pharmaceutical company GlaxoSmithKline, to develop the malaria vaccine RTS,S in the late 1980s. RTS,S has progressed further than any other therapy for the disease, through phase III clinical trials. This 'subunit vaccine' has been shown to reduce clinical infection in 5- to 17-month-old babies (M. T. White *et al. Lancet Infect. Dis.* **15**, 1450–1458; 2015). It should be rolled out in the next year for further testing in Ghana, Kenya and Malawi, on the basis of recommendations by the WHO.

A third vaccine strategy was also inspired by a major finding made by Nussenzweig's research group. Her team demonstrated that $CD8^+$ T cells, a type of white blood cell that mediates protection against viruses and tumours, have an important role in how attenuated sporozoites confer immunity in the liver (P. Romero *et al. Nature* **341**, 323–326; 1989). This has led researchers to investigate vaccines that induce T-cell immunity to kill parasites in the organ.

Nussenzweig was the first woman to chair a department at the NYU School of Medicine. Under her leadership, and in collaboration with Victor, the department trained hundreds of students and fellows — many came from Brazil or later went to work there. Starting in 2012, she and Victor themselves returned to do research in Brazil for some months each year, at the Federal University of São Paulo (UNIFESP). Many of her disciples are now leaders in the field. She and Victor also inspired a love of science in their three children, all of whom are successful academics.

A defiance of established paradigms and a relentless pursuit of new ideas defined Nussenzweig as a scientist. She was a devoted experimentalist, which give her work a solidity that allowed once-controversial ideas to become mainstream. She was always excited when discussing research, and was a visionary role model, mentor and colleague.

Respect for her legacy requires us to match her determination and focus, to finally bring an end to malaria. ■

Robert A. Seder *is at the National Institute for Allergy and Infectious Diseases in Bethesda, Maryland; he met Ruth in 2005 through his work on vaccines, specifically, the PfSPZ vaccine.* **Fidel Zavala** *is at the Johns Hopkins Bloomberg School of Public Health in Baltimore, Maryland, and is a member of the Malaria Research Institute. He met Ruth in 1980 as a postdoctoral fellow; they worked together for 23 years.*
e-mail: fzavala1@jhu.edu

ASTRONOMY

Early star formation detected

Little is known about the star-birth activity of the earliest galaxies. Observations of a particularly distant galaxy provide evidence for such activity when the Universe was just 2% of its current age. SEE LETTER P.392

RYCHARD BOUWENS

Astronomers have only a limited understanding of star and galaxy formation in the first 300 million years of the Universe's history[1,2]. Two important processes would have been accretion, whereby objects increase in mass by gravitationally attracting nearby matter, and the cooling of gas. But the overall picture lacks the sensitive observational data required to inform or guide its construction. On page 392, Hashimoto *et al.*[3] present observations of an extremely distant galaxy, and report that star-birth activity began there just 250 million years after the Big Bang. The authors' results suggest that future telescopes could detect such early episodes of star formation in similar galaxies.

The first stars in the Universe are thought to have formed in regions with high densities of matter[1,2]. These regions grew over cosmic time by accretion, and eventually developed into galaxies. From a theoretical standpoint, there is great interest in establishing observationally when the Universe had its first major star-birth activity, but current constraints are poor. One way in which to investigate the onset of star formation is to detect starlight from extremely distant galaxies. Because light travels at a finite speed, observations of the distant Universe act as a time machine, allowing us to look back into the past.

The galaxy probed by Hashimoto and colleagues is one of the farthest known objects from Earth for which light can be detected (Fig. 1a). It was discovered[4] in 2012, and was dubbed MACS1149-JD1. What the authors have added to this discovery is a precise measurement of the galaxy's redshift. The redshift of a light source tells us the factor by which the Universe has expanded since the source emitted its light, as well as the distance to the source and the time at which the light was released.

Hashimoto *et al.* determined the redshift of MACS1149-JD1 by studying the properties of an emission line in the galaxy's spectrum. They report a redshift of about 9.11, which implies that the galaxy is being viewed as it was when the Universe was roughly one-tenth of its current size and about 550 million years old. This is the highest redshift ever inferred from a spectral line[5], and is only slightly lower than a redshift reported by using a broader, possibly less robust, spectral feature[6].

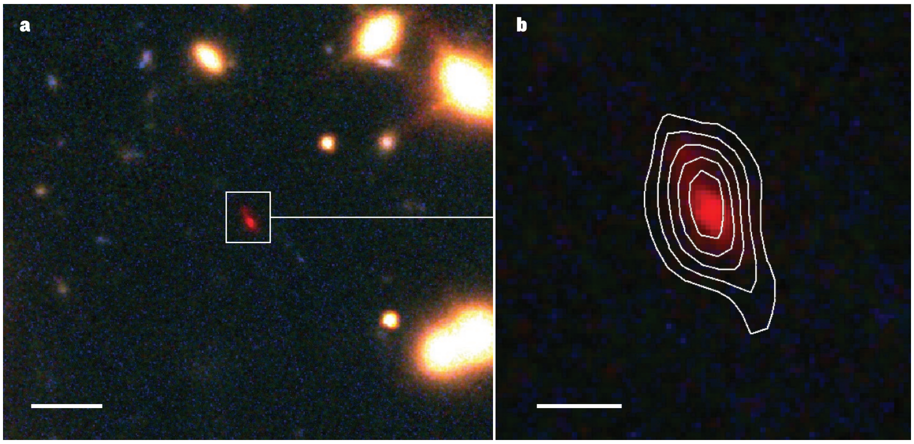


Figure 1 | Observations of MACS1149-JD1. a, Hashimoto *et al.*[3] report that star formation in the distant galaxy MACS1149-JD1 (indicated by the box) began when the Universe was only 2% of its current age. Other sources of light in the image are nearby galaxies. Scale bar, 2 arcseconds (one arcsec is 1/3,600th of a degree). **b**, The authors' result relied on a detailed study of an emission line in the galaxy's spectrum that is produced by doubly ionized oxygen. The contours join points that have equal emission-line intensity. Scale bar, 0.5 arcsec. (Adapted from Extended Data Fig. 2a of ref. 3.)

The authors' measurement was made possible thanks to the continuously improving capabilities of the US$1.4-billion Atacama Large Millimeter/submillimeter Array (ALMA) observatory in Chile. ALMA studies of galaxies with redshifts greater than 4 frequently use an emission line that is produced by singly ionized carbon[7–9]. By contrast, Hashimoto and colleagues relied on a line that is associated with doubly ionized oxygen (Fig. 1b). This line was shown to be readily detectable in previous theoretical and observational work using ALMA[10,11]. The authors' result showcases ALMA's capabilities as a tool for precisely measuring the redshifts of distant galaxies, as was also illustrated earlier this year[7].

Next, Hashimoto *et al.* considered the optical colours observed in MACS1149-JD1 by NASA's Hubble and Spitzer space telescopes. Such colours provide clues about the number of stars that formed early in a galaxy's lifetime[4,12,13]. The authors show that the colours represent a substantial episode of star formation in the galaxy when the Universe was only 250 million years old. The authors' precise redshift measurement was essential in arriving at this conclusion because it allowed them to rule out the possibility that the colours arise instead from strong recombination lines — spectral features that are commonly associated with the intense ionizing radiation produced by hot stars in the earliest galaxies[14–16].

However, MACS1149-JD1 is just one galaxy, and it remains unclear whether such early star-birth activity occurred in other galaxies. Observations of the cosmic microwave background — the relic radiation from the Big Bang — by the Planck Collaboration[17] indicate that there was less star formation in the early Universe than the present authors' results suggest. In addition, measurements of the prevalence of galaxies at similar epochs to that of MACS1149-JD1 suggest a lack of star-birth activity at these early times[18]. Nevertheless, a high rate of star formation in the early Universe could explain the discovery earlier this year of an unexpectedly large absorption signal in the spectrum of the cosmic microwave background[19].

Hashimoto and colleagues have not only set a record for the highest redshift inferred from a spectral line, but they also did it using ALMA, which is a first for the facility. The possibility that the galaxy they observed had substantial star formation at early times is intriguing. Their discoveries seem certain to inspire similar studies of other galaxies in the distant Universe, and provide fuel for observations using the future James Webb Space Telescope. ■

Rychard Bouwens *is at the Leiden Observatory, Leiden University, Leiden NL-2333, the Netherlands.*
e-mail: bouwens@strw.leidenuniv.nl

1. Loeb, A. & Barkana, R. *Annu. Rev. Astron. Astrophys.* **39**, 19–66 (2001).
2. Abel, T., Bryan, G. L. & Norman, M. L. *Science* **295**, 93–98 (2002).
3. Hashimoto, T. *et al. Nature* **557**, 392–395 (2018).
4. Zheng, W. *et al. Nature* **489**, 406–408 (2012).
5. Zitrin, A. *et al. Astrophys. J. Lett.* **810**, 12 (2015).
6. Oesch, P. A. *et al. Astrophys. J.* **819**, 129 (2016).
7. Smit, R. *et al. Nature* **553**, 178–181 (2018).
8. Capak, P. L. *et al. Nature* **522**, 455–458 (2015).
9. Willott, C. J., Carilli, C. L., Wagg, J. & Wang, R. *Astrophys. J.* **807**, 180 (2015).
10. Inoue, A. K. *et al. Astrophys. J. Lett.* **780**, 18 (2014).
11. Inoue, A. K. *et al. Science* **352**, 1559–1562 (2016).
12. Bradač, M. *et al. Astrophys. J.* **785**, 108 (2014).
13. Hoag, A. *et al. Astrophys. J.* **854**, 39 (2018).
14. Roberts-Borsani, G. W. *et al. Astrophys. J.* **823**, 143 (2016).
15. Smit, R. *et al. Astrophys. J.* **784**, 58 (2014).
16. Laporte, N. *et al. Astron. Astrophys.* **562**, 8 (2014).
17. Planck Collaboration. *Astron. Astrophys.* **596**, 107 (2016).
18. Oesch, P. A., Bouwens, R. J., Illingworth, G. D., Labbé, I. & Stefanon, M. *Astrophys. J.* **855**, 105 (2018).
19. Bowman, J. D., Rogers, A. E. E., Monsalve, R. A., Mozdzen, T. J. & Mahesh, N. *Nature* **555**, 67–70 (2018).

COMPUTATIONAL NEUROSCIENCE

AI mimics brain codes for navigation

An artificial-intelligence technique called deep learning has now been used to model spatial navigation. The system develops a representation of space similar to that of the grid cells found in the mammalian brain. See Letter p.429

FRANCESCO SAVELLI & JAMES J. KNIERIM

Deep learning is an approach to artificial intelligence that is inspired by the brain's neural networks. The technique is contributing to a plethora of technologies, from automated video analysis to language translation. One page 429, Banino *et al.*[1] use this framework to gain insights into real-life neuronal networks — in particular, how geometrically regular representations of space can facilitate flexible navigation strategies.

Deep-learning networks can be taught how to process inputs to achieve a particular output — for instance, learning to pick out a particular face in many photos of different people. The networks are 'deep' in that they are made up of sequential layers of repeated computational units. Each unit receives inputs from similar units in the previous layer and sends outputs to those in the next. Mathematically, such a network can be viewed as a high-dimensional function, which can be modulated by altering how the outputs of one layer are weighted in the next.

The network tunes the function during a training phase, which typically relies on a set of input–output examples. For instance, a deep-learning system might be shown a series of photos, and told which ones contain the face it aims to identify. Its weights are automatically tuned by optimization algorithms until it learns to make a correct identification. The network's deep organization gives it a prodigious ability to spot and take advantage of the most useful features and patterns that recur across the examples, and distinguish different faces. But one downside is that the final network tends to be a black box — the computational solutions derived during training often cannot be deciphered from the myriad weights assigned throughout the layers.

Deep-learning networks can successfully perform perceptual tasks[2], but there have been fewer studies of complex behavioural tasks such as navigation. A key aspect of real-life navigation is estimating one's position following each step, by calculating the displacement per step on the basis of orientation and distance travelled. This process is called path integration and is thought by neuroscientists, cognitive scientists and roboticists to be crucial for generating a cognitive map of the environment[3–5]. There are several kinds of neuron associated with the brain's cognitive maps, including place cells, which fire when the organism occupies a particular position in the environment, and head-direction cells, which signal head orientation.

A third type of neuron, the grid cell, fires when the animal is at any of a set of points that form a hexagonal grid pattern across the environment. Grid cells are thought to endow the cognitive map with geometric properties that help in planning and following trajectories. These cells are found in the brain's hippocampal formation, a region that, in humans, is involved in spatial learning, autobiographical memories and knowledge of general facts about the world.

Banino and colleagues set out to generate path integration in a deep-learning network. Because path integration involves

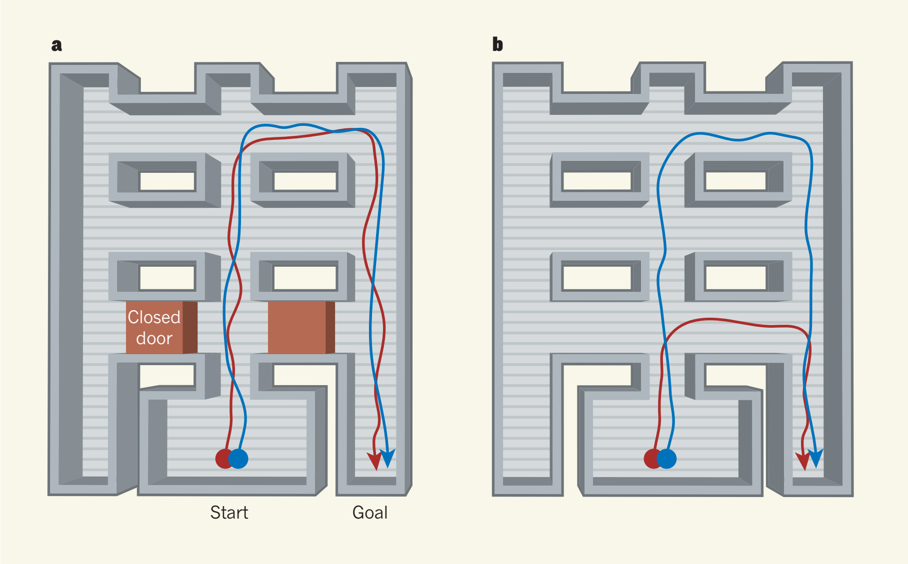


Figure 1 | An AI system learns to take shortcuts. In the mammalian brain, place cells fire when an animal is at a particular position within an environment, head-direction cells fire when the head is in a particular orientation, and grid cells fire when the animal is at points that form a hexagonal grid across the environment. Banino and colleagues[1] trained an artificial-intelligence system called a deep-learning network to navigate, by providing it with simulations of rodent foraging patterns, including about the activity of place and head-direction cells. Some computational units in the network developed grid-cell-like firing patterns (not shown). **a**, While learning to navigate towards a goal, similar paths were taken by both a system using grid cells (red line indicates a sample path) and a system that used place and head-direction cells instead (blue line). **b**, But when shortcuts were introduced, for example by opening previously closed doors, only the system that used grid cells found the shorter routes, highlighting the ability of grid-cell-like activity to promote flexible navigation strategies. (Figure adapted from Extended Data Fig. 10 in ref. 1.)

remembering the output from the previous processing step and using it as input for the next, the authors used a network involving feedback loops. They trained the network using simulations of pathways taken by foraging rodents. The system received information about the simulated rodent's linear and angular velocity, and about the simulated activity of place and head-direction cells — the latter two acting as an 'oracle' for the current location and head direction of the rodent.

The authors found that patterns of activity resembling grid cells spontaneously emerged in computational units in an intermediate layer of the network during training, even though nothing in the network or the training protocol explicitly imposed this type of pattern. The emergence of grid-like units is an impressive example of deep learning doing what it does best: inventing an original, often unpredicted internal representation to help solve a task.

Grid-like units allow the network to keep track of position on the basis of path integration. Can they also help the system to learn to navigate efficiently from its current position to a goal location? To address this question, Banino *et al.* added a reinforcement-learning component, in which the network learned to assign values to specific actions taken at specific locations. Higher values were assigned to actions that brought the simulated rodent closer to the goal, acting as a reward. The grid-like representation markedly improved the ability of the network to solve goal-directed tasks, compared to control simulations in which the start and goal locations were encoded instead by place and head-direction cells. The trained network found smarter shortcuts when obstacles such as closed doors were removed (Fig. 1), and even extrapolated paths towards goals in a previously unexplored annex of a familiar environment. These results support the idea that grid cells enable the brain to perform vector calculations (calculations about the length and direction of a path) to assist path planning through compartmentalized[6] or previously unexplored[7] environments.

In the future, the authors' network could be used to explore the consequences of interactions between grid and place cells. In the current network, the simulated place layer does not change during training. However, in the brain, grid and place cells influence each other in ways that are not well understood. Although real-life place cells can remain spatially selective in the absence of grid-cell inputs[8], these inputs seem important when an animal is far from external landmarks that can be used to define locations[9–11]. Under these conditions, place cells presumably rely on path integration and grid cells to maintain an accurate estimate of position. By developing the network such that the place-cell layer can be modulated by grid-like inputs, we could begin to unpack this relationship.

From a broader perspective, it is interesting that the network, starting from very general computational assumptions that do not take into account specific biological mechanisms, found a solution to path integration that seems similar to the brain's. That the network converged on such a solution is compelling evidence that there is something special about grid cells' activity patterns that supports path integration. The black-box character of deep-learning systems, however, means that it might be hard to determine what that something is.

Likewise, the fact that the grid representation enhanced goal-directed performance is a compelling proof-of-concept of the role of grid cells in the brain. But the authors had to use correlational analyses, guided by qualitative intuitions, to indirectly infer that the network was making vector calculations. The inability to directly manipulate these calculations in the model makes it difficult to examine the computational principles, algorithms and encoding strategies that make grid-cell representations of space such an efficient solution for navigation. As such, the theoretician ends up in the same quandary as the experimentalist: trying to tease apart a poorly understood complex system to understand it. Making deep-learning systems more intelligible to human reasoning is an exciting challenge for the future. ■

Francesco Savelli *and* **James J. Knierim** *are at the Zanvyl Krieger Mind/Brain Institute, Johns Hopkins University, Baltimore, Maryland 21218, USA.* **J.J.K.** *is also in the Solomon H. Snyder Department of Neuroscience, Johns Hopkins University.*
e-mails: fsavelli.research@gmail.com; jknierim@jhu.edu

1. Banino, A. *et al. Nature* **557**, 429–433 (2018).
2. LeCun, Y., Bengio, Y. & Hinton, G. *Nature* **521**, 436–444 (2015).
3. McNaughton, B. L., Battaglia, F. P., Jensen, O., Moser, E. I. & Moser, M.-B. *Nature Rev. Neurosci.* **7**, 663–678 (2006).
4. Gallistel, C. R. *The Organization of Learning* (MIT Press, 1990).
5. Thrun, S., Burgard, W. & Fox, D. *Probabilistic Robotics* (MIT Press, 2005).
6. Carpenter, F., Manson, D., Jeffery, K., Burgess, N. & Barry, C. *Curr. Biol.* **25**, 1176–1182 (2015).
7. Savelli, F., Luck, J. D. & Knierim, J. J. *eLife* **6**, e21354 (2017).
8. Koenig, J., Linder, A. N., Leutgeb, J. K. & Leutgeb, S. *Science* **332**, 592–595 (2011).
9. Muessig, L., Hauser, J,. Wills, T. J. & Cacucci, F. *Neuron* **86**, 1167–1173 (2015).
10. Wang, Y., Romani, S., Lustig, B., Leonardo, A. & Pastalkova, E. *Nature Neurosci.* **18**, 282–288 (2015).
11. Mallory, C. S., Hardcastle, K., Bant, J. S. & Giocomo, L. M. *Nature Neurosci.* **21**, 270–82 (2018).

OCEANOGRAPHY

Climate change and oxygen in the ocean

Computer simulations show that areas of the ocean that have low levels of dissolved oxygen will expand, but then shrink, in response to global warming — adding to an emerging picture of the finely balanced processes involved.

LAURE RESPLANDY

Global warming has reduced the amount of dissolved oxygen in the ocean by 2% since 1960 (ref. 1). A major concern is that the rate of loss of dissolved oxygen has already increased by up to 20% in tropical waters, expanding the volume of regions called oxygen minimum zones (OMZs), where levels of dissolved oxygen are already very low[2,3]. The expansion of tropical OMZs threatens the survival of marine organisms that rely on dissolved oxygen for respiration, and affects the biogeochemical cycling of carbon and nitrogen, potentially amplifying global warming[4]. Writing in *Global Biogeochemical Cycles*, Fu *et al.*[5] suggest that, in the long term, tropical OMZs might shrink after their initial expansion, reversing their impact on warming.

Although, overall, tropical oceans have lost dissolved oxygen in the past 50 years, observations indicate strong regional and temporal variations. For example, OMZs have clearly expanded since the 1970s in the equatorial Pacific and Atlantic oceans[3,6], but a long-term reconstruction[7] of the OMZ close to the Californian and Mexican coasts suggests that it had been shrinking for a century before it started to expand in the 1990s. In the Indian Ocean, the northern part of the OMZ is shrinking, whereas the southern part is expanding[8].

Three competing processes control the levels of dissolved oxygen in the ocean. The first is the transfer of atmospheric oxygen to the surface ocean (which, in turn, is tied to oxygen solubility, the capacity of the water to hold on to dissolved oxygen). The second is ocean circulation (ventilation), which carries the oxygen-rich surface waters to the ocean interior. And, finally, there is the biological respiration of dissolved oxygen, the process by which microorganisms consume organic matter produced at the surface as it sinks to the deep ocean. OMZs develop in tropical intermediate waters (those at depths of 200–1,000 metres) because the physical supply of oxygen is low, but the biological demand for it is high.

As the ocean warms, the solubility of oxygen within it declines, and vertical stratification — the formation of distinct layers of water — increases, limiting the exchange of oxygen between surface and intermediate waters. Together, these changes reduce the supply of oxygen to the ocean interior and expand OMZs. Increased stratification also reduces the supply of nutrients to surface waters, limits biomass production[9] and reduces the export of organic matter to the deep ocean. This biological effect lowers the demand for oxygen in underlying waters and reduces the volume of OMZs. In addition, local changes in ventilation in OMZs can act to shrink these zones[5,10].

Globally, the direct effect of declining solubility explains half of the observed loss of oxygen in the upper 1,000 m of the ocean, and a combination of biological and ventilation effects accounts for the other half[1]. In tropical OMZs, however, the balance between these processes is more subtle and the relative contributions harder to tease out[11]. Numerical models of Earth systems can provide a mechanistic understanding of these past changes in oxygen and why they vary in space, and can inform us about changes to expect in the future.

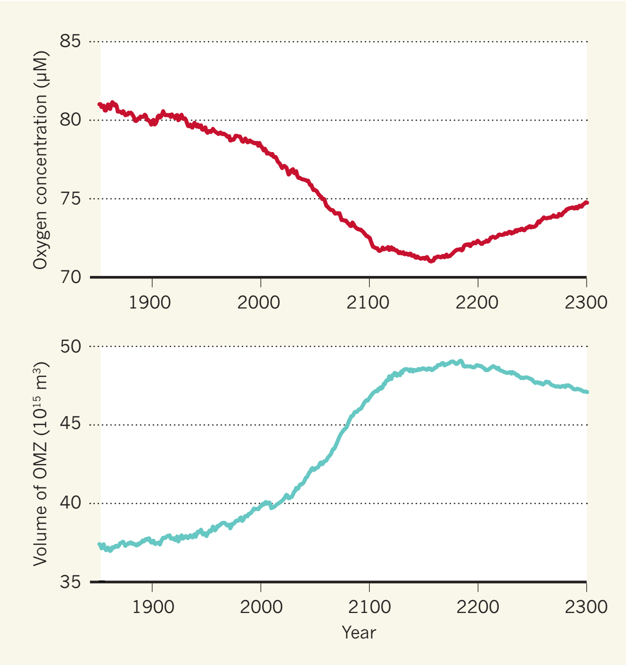


Figure 1 | Projections of oxygen concentrations and volumes of oxygen minimum zones (OMZs)in tropical seas. **a**, Fu *et al.*[5] have used computer simulations of the Earth system to reproduce the observed historical decline in dissolved-oxygen concentrations (measured in micromoles per litre) in tropical-ocean intermediate waters (which occur at latitudes between 30° N and 30° S at depths of 200–1,000 metres), and to project how such oxygen levels might vary between the twenty-first and twenty-third centuries. The simulations suggest that oxygen levels will decline until about 2150, level off and then increase until 2300. **b**, The authors also simulate changes in the volume of oxygen minimum zones — ocean regions where dissolved-oxygen levels are less than 20 μM. The volume mirrors the pattern shown in **a**. However, the factors that contribute to the simulated changes are finely balanced.

Fu and colleagues used long-term simulations of the Earth system to reproduce the observed historical decline in dissolved-oxygen levels in the tropical ocean, and to project the evolution of OMZs between the twenty-first and twenty-third centuries (Fig. 1). In their model, the volume of OMZs initially expands, then levels off around the year 2150, and finally shrinks until 2300. The authors also introduce a modelling framework that can disentangle the contributions to this pattern that are made by solubility, ventilation and the reduction in biological export and respiration. In their model, the shift from OMZ expansion to contraction occurs when the biological effects become more prominent than the solubility effect, in 2150. Ventilation further tips this balance, as it switches from an initial slowdown that reinforces the expansion effect of solubility before 2150, to a reinvigoration that reinforces the OMZ contraction controlled by biological effects after 2150.

The findings suggest that about half of the OMZ expansion that occurred between 1900 and 2150 could be reversed by the year 2300. Moreover, the core of the OMZ (the region that has the lowest levels of dissolved oxygen) might contract to become smaller than it was in pre-industrial times, and could possibly act as a negative climate feedback that dampens global warming. The formation of a smaller OMZ core would certainly limit the production of nitrous oxide[12], a powerful greenhouse gas, and limit the consumption of nitrate nutrients for respiration, which in turn could boost the efficiency of the ocean's biological carbon sink[13].

Fu and co-workers' study highlights the extreme sensitivity of OMZs to biological and physical changes, and the difficulty of quantifying and anticipating these changes. Different Earth-system models currently project dramatically different changes in OMZ volume (ranging from −2% to +16% by 2100)[14]. So how much confidence can we have in OMZ projections?

In fact, Earth-system models agree on many aspects of future OMZs. Modelled changes in oxygen solubility are relatively well constrained and are tied to surface warming[14,15]. The reduction in biological export identified by Fu *et al.* is robustly produced by other models[11,14], and most models also project a strengthening of the ventilation in tropical OMZs[16]. Uncertainties arise from the differences in the magnitude and timing of these changes. Simulated reductions in biological export by 2100 vary between 1% and 40%[15], and changes in ventilation can vary by a factor of two between models[16]. These differences, even when small, can tip the balance of oxygen levels in OMZs, shifting them from expansion to contraction.

The challenge now is to constrain this subtle balance and the possible shifts in OMZ volume not only in time, as Fu *et al.* have done, but also between regions. For example, reduced biological export and consumption of oxygen is likely to have driven the OMZ contraction reported along the Californian and Mexican coasts[7], whereas weakening ventilation probably underlies the OMZ expansion in the equatorial Pacific. Other factors, such as the supply of nutrients from anthropogenic aerosols, winds or ocean acidification, could also locally offset the effects of changes in solubility and stratification[17]. OMZ climate feedbacks — the reinforcing or dampening of climate change by OMZs — are tied to global OMZ volume, but establishing local OMZ responses will be crucial if we are to assess the impacts on ecosystems and ecosystem services, such as fisheries. ■

Laure Resplandy *is in the Department of Geosciences, Princeton Environmental Institute, Princeton University, Princeton, New Jersey 085443, USA.*
e-mail: laurer@princeton.edu

1. Schmidtko, S., Stramma, L. & Visbeck, M. *Nature* **542**, 335–339 (2017).
2. Levin, L. A. *Annu. Rev. Mar. Sci.* **10**, 229–260 (2018).
3. Stramma, L., Johnson, G. C., Sprintall, J. & Mohrholz, V. *Science* **320**, 655–658 (2008).
4. Keeling, R. F., Körtzinger, A. & Gruber, N. *Annu. Rev. Mar. Sci.* **2**, 199–229 (2010).
5. Fu, W., Primeau, F., Moore, J. K., Lindsay, K. & Randerson, J. T. *Glob. Biogeochem. Cycles* **32**, https://doi.org/10.1002/2017GB005788 (2018).
6. Brandt, P. *et al. J. Phys. Oceanogr.* **40**, 1784–1801 (2010).
7. Deutsch, C. *et al. Science* **345**, 665–668 (2014).
8. Banse, K., Naqvi, S. W. A., Narvekar, P. V., Postel, J. R. & Jayakumar, D. A. *Biogeosciences* **11**, 2237–2261 (2014).
9. Behrenfeld, M. J. *Nature Clim. Change* **4**, 880–887 (2014).
10. Gnanadesikan, A., Russell, J. L. & Zeng, F. *Ocean Sci.* **3**, 43–53 (2007).
11. Cabré, A., Marinov, I., Bernardello, R. & Bianchi, D. *Biogeosciences* **12**, 5429–5454 (2015).
12. Babbin, A. R., Bianchi, D., Jayakumar, A. & Ward, B. B. *Science* **348**, 1127–1129 (2015).
13. Gruber, N. in *The Ocean Carbon Cycle and Climate* (eds Follows, M. & Oguz, T.) 97–148 (Springer, 2004).
14. Bopp, L. *et al. Biogeosciences* **10**, 6225–6245 (2013).
15. Bopp, L., Resplandy, L., Untersee, A., Le Mezo, P. & Kageyama, M. *Phil. Trans. R. Soc. A* **375**, 20160323 (2017).
16. Shigemitsu, M., Yamamoto, A., Oka, A. & Yamanaka, Y. *Glob. Biogeochem. Cycles* **31**, 804–820 (2017).
17. Oschlies, A. *et al. Phil. Trans. R. Soc. A.* **375**, 20160325 (2017).

CELL REPROGRAMMING

More than one way to induce a neuron

Seventy-six pairs of transcription factors can induce mouse connective-tissue cells to adopt a neuron-like identity *in vitro*. This discovery provides insights into both neuronal development and cell reprogramming. SEE ARTICLE P.375

LYNETTE LIM & OSCAR MARÍN

The brain contains hundreds of neuronal subtypes, each defined by a specific combination of features, including its position and shape, the neurotransmitter molecules it produces and its electrophysiological properties[1]. Engineering this enormous diversity in the laboratory is an ultimate goal of regenerative medicine. On page 375, Tsunemoto *et al.*[2] describe a large-scale effort to identify factors that can endow non-neural cells cultured *in vitro* with neuronal properties.

Understanding the mechanisms that underlie the generation of neuronal diversity has been a central goal of neurobiology for more than a hundred years, since the neuroscientist Santiago Ramón y Cajal postulated that the nervous system is made up of discrete individual cells[3]. Work over the past three decades has identified gene regulatory networks that control neuronal identity as it unfolds in the embryonic brain[4]. These studies have also revealed that neuronal identity is intimately linked to the environment in which neurons develop, primarily because some of the cells' most important features, such as their connections, depend on their interactions with other neurons.

In the past decade, however, it has become clear that many neuronal attributes can be generated outside the normal context of brain development. For example, in 2010 it emerged[5] that a cocktail of three transcription factors can be applied to fibroblasts (the most common cells of connective tissue) cultured *in vitro* to convert them into cells that resemble brain-derived neurons, at least in terms of their shape and electrophysiological properties. This procedure, called direct lineage reprogramming, is based on the premise that certain transcription factors regulate gene-expression patterns characteristic of neuronal cell types. But what has not been clear is whether the capacity to reprogram cells into neurons is limited to a handful of transcription factors.

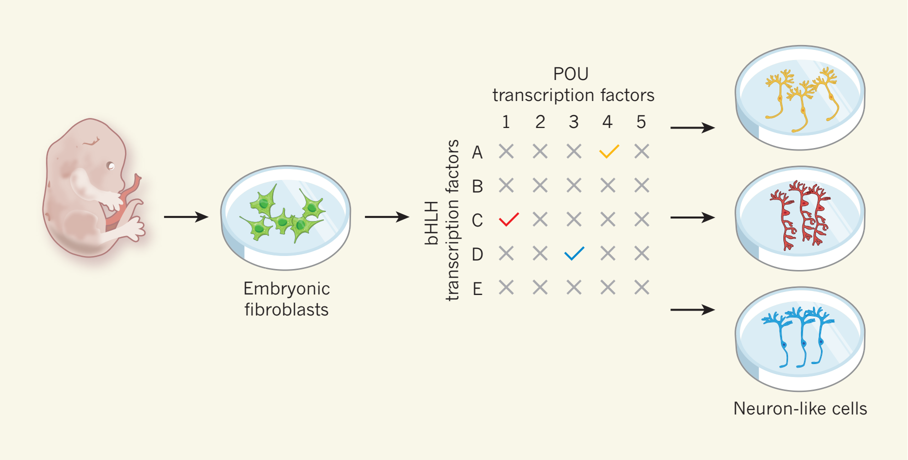
Previous work by the group that performed the current study showed that a pair of transcription factors from the basic helix-loop-helix (bHLH) and Pit-Oct-Unc (POU) families can induce the expression of neuronal markers through direct reprogramming[6]. Tsunemoto *et al.* were inspired by this finding in their current work. The authors screened 598 pairs of bHLH and POU transcription factors — chosen on the basis of their expression in neuronal lineages — to see which could transform mouse embryonic fibroblasts into neurons *in vitro* (Fig. 1). Seventy-six of the pairs produced cells that expressed multiple markers of mature neurons and had neuronal morphologies. Thus, neuronal features can be induced in non-neuronal cells by an astonishing range of transcription-factor combinations.

How similar are the neurons induced by the transcription-factor pairs? Analysis of gene expression using single-cell RNA sequencing revealed that a given pair of factors generates relatively homogeneous populations of neurons, which share a similar molecular profile. This is surprising, because previous experiments have highlighted the heterogeneity of cell populations undergoing reprogramming in culture[7].

By contrast, Tsunemoto *et al.* found that different transcription-factor combinations induced the formation of distinct neuron-like populations that had characteristic markers and electrophysiological features. However, the authors also found that certain features — such as the expression of particular ion channels or neurotransmitter receptors — could be induced by multiple combinations of transcription factors. These findings support the idea that there is not a single 'gene code' for making a particular feature of neuronal identity, but that the molecular machinery underlying neuronal development has a remarkable degree of redundancy. In other words, identity-defining transcription factors are probably used in different combinations in distinct neuronal cell types to generate the diversity seen *in vivo*.

One key question is whether the induced neurons faithfully mimic neuronal cell types found *in vivo*, or whether (and to what extent) they represent artificial cell types. To address this issue, the authors compared gene-expression patterns in induced cell populations with those in neuronal subtypes taken from three-week-old mice. This analysis indicated that the induced neurons did not match endogenous cell types of the juvenile mouse brain. However, it might be that the induced neurons did not reach the same stage of development as endogenous cells. Matching cell types across different developmental stages remains a complex challenge in neurobiology, as shown this year by two RNA-sequencing studies[8,9]. These highlighted the difficulty of recognizing the transcription factors that define specific cell types at early stages of neuronal development in the brain's cerebral cortex.

It is possible, then, that the neurons induced *in vitro* by Tsunemoto *et al.* correspond to specific populations of endogenous neurons in a relatively immature state. These cells might develop into fully differentiated neurons only if placed in the appropriate environment. Alternatively, the expression of a pair of transcription factors might be sufficient to elicit the development of some neuronal features in fibroblasts, but not to unleash the complete program of



	1	2	3	4	5
A	×	×	×	✓	×
B	×	×	×	×	×
C	✓	×	×	×	×
D	×	×	✓	×	×
E	×	×	×	×	×

Figure 1 | Inducing neurons through direct reprogramming. Tsunemoto *et al.*[2] grew embryonic fibroblasts (connective-tissue cells isolated from mouse embryos) *in vitro*. They treated the cells with 598 pairs of transcription factors from the basic helix-loop-helix (bHLH) and Pit-Oct-Unc (POU) families, most of which are expressed in neuronal lineages. In total, more than 12% of the transcription-factor combinations tested could reprogram the fibroblasts into cells with neuronal properties (indicated with ticks). Different pairs produced neuron-like cells with different shapes, gene-expression profiles and electrophysiological properties (indicated by different colours).

differentiation that takes place in the embryo.

Nevertheless, Tsunemoto and colleagues' study adds to the body of evidence showing that some features of neuronal identity can be reproduced outside the developing brain. In doing so, it demonstrates the power of reprogramming to interrogate the function of neuron-specific genes. The authors have made their findings available in a database (see go.nature.com/2r1msxi) that will allow other researchers to use the transcription-factor codes to induce specific neuronal features. This will doubtless prove useful for studying the selective vulnerability of specific neuronal subtypes to disease.

Finally, the authors provide preliminary evidence that their transcription-factor combinations can also be used to generate neurons from human embryonic fibroblast-like cells. Following further validation, the codes might help us to decipher the origins of neuronal diversity in humans. ■

Lynette Lim *and* **Oscar Marín** *are in the Centre for Developmental Neurobiology, Institute of Psychiatry, Psychology and Neuroscience, King's College London, London SE1 1UL, UK, and in the MRC Centre for Neurodevelopmental Disorders, King's College London.*

e-mail: oscar.marin@kcl.ac.uk

1. Zeng, H. & Sanes, J. R. *Nature Rev. Neurosci.* **18**, 530–546 (2017).
2. Tsunemoto, R. *et al. Nature* **557**, 375–380 (2018).
3. Ramón y Cajal, S. *Histologie du Système Nerveux de l'Homme et des Vertébrés* (Maloine, 1909).
4. Lodato, S. & Arlotta, P. *Annu. Rev. Cell. Dev. Biol.* **31**, 699–720 (2015).
5. Vierbuchen, T. *et al. Nature* **463**, 1035–1041 (2010).
6. Blanchard, J. W. *et al. Nature Neurosci.* **18**, 25–35 (2015).
7. Treutlein, B. *et al. Nature* **534**, 391–395 (2016).
8. Mi, D. *et al. Science* **360**, 81–85 (2018).
9. Mayer, C. *et al. Nature* **555**, 457–462 (2018).

ATMOSPHERIC SCIENCE

Increased emissions of ozone depleters

Chlorofluorocarbons are the main class of chemical that depleted the ozone layer in the stratosphere. Measurements reveal that emissions of these compounds are rising again, despite international rules restricting their use. SEE LETTER P.413

MICHAELA I. HEGGLIN

Monitoring the expected decline in the atmospheric concentrations of banned compounds might seem like an unexciting research task. But on page 413, Montzka *et al.*[1] report an unexpected finding in the long-term measurements of CFC-11, one of the most potent ozone-depleting compounds: its atmospheric concentration is decreasing much more slowly than would be expected on the basis of its known sources and sinks. This points to a fresh rise in emissions — in contravention of international regulations.

CFC-11 belongs to the chlorofluorocarbon (CFC) family of compounds. CFCs are highly stable, synthetic chemicals that were used in various applications from the 1930s onwards — for example, as propellants in aerosol sprays, solvents and refrigerants. In the early 1970s, the British chemist James Lovelock and his colleagues were the first to measure the abundance of CFCs in the atmosphere and to realize that these substances were found ubiquitously in both the Northern and Southern hemispheres, despite their sources being located only in the Northern Hemisphere[2]. This finding led to the hypothesis that CFCs could be destroyed naturally only in the stratosphere, in a process that releases chlorine atoms. Each of these atoms would be able to destroy many ozone molecules in catalytically driven cycles, thus posing a threat to the ozone layer[3], which protects life on Earth from harmful ultraviolet radiation.

The discovery[4] of the 'hole' in the ozone layer over Antarctica in 1985 proved this hypothesis to be not only correct, but also much more threatening than had been imagined. It spurred research activities to understand why such severe ozone depletion was found over Antarctica alone, and led to political action to restrict the use of CFCs under the Montreal Protocol in 1987. The realization that more-severe ozone depletion would spread further across the globe if CFCs continued to be released into the atmosphere, along with technological advances that made the replacement of CFCs possible, helped governments to tighten the regulations on CFCs and ultimately ban their production through several amendments to the protocol. As a result of these actions, CFC concentrations in the atmosphere peaked in the mid-to-late 1990s and have been steadily declining ever since[5]. The Montreal Protocol has been hailed as the most successful international treaty so far that deals with a global environmental issue[6].

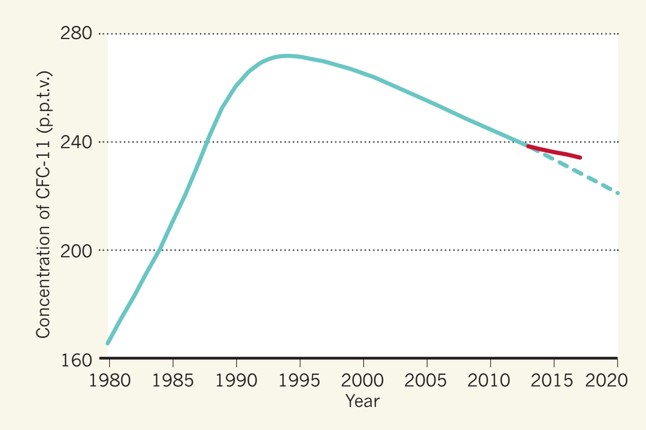


Figure 1 | The slowing decline of an ozone-depleting compound. Production of chlorofluorocarbons (CFCs) was banned internationally by the Montreal Protocol in 1987 because CFC emissions lead to destruction of the stratospheric ozone layer. The blue line shows the atmospheric abundance[5] of one such compound, CFC-11, in parts per trillion by volume (p.p.t.v.); the solid part indicates measurements up to 2013, and the dashed part indicates the projected abundance, assuming that no new CFC-11 is produced. Montzka *et al.*[1] report that the atmospheric abundance of CFC-11 over the past few years (red line; approximated from Fig. 1 of the paper[1]) is greater than had been projected, and so the rate of decline of CFC-11 levels is slower than expected. They conclude that CFC-11 emissions must have increased since 2012.

Because the destruction of CFCs in the stratosphere is a slow process, their removal from the atmosphere will take many decades. Today's research into stratospheric ozone focuses on whether atmospheric concentrations of ozone-depleting substances are decreasing as expected, and whether the ozone layer is on its way to recovery. Working out whether the ozone layer is recovering on the basis of ozone observations alone is particularly difficult because of confounding effects from natural variability, climate change and ozone pollution[7].

Monitoring the atmospheric concentration of ozone-depleting substances such as CFC-11 is a more direct test of the effectiveness of the Montreal Protocol. However, even for these long-lived chemicals, natural variability in the transport of air masses between the source and sink regions of the chemicals can affect the rate of the expected decline. The source regions are mostly found in highly industrialized areas of the Northern Hemisphere, whereas CFC destruction in the stratosphere (in both hemispheres) acts as the sink. This distribution of sources and sinks leads to concentration differences between the Northern and Southern hemispheres, which get smaller over time after emissions cease. The rate of exchange of air between the stratosphere and the underlying troposphere, and between the Northern and Southern hemispheres, both have a crucial role in driving the concentrations of ozone-depleting substances.

Montzka *et al.* have made a rigorous attempt to take into account natural variability in transport between the different regions of the atmosphere to calculate how it might have generated the observed levels of CFC-11, both by using simple ‘box’ models and by performing 3D computational simulations using comprehensive climate models that take into account atmospheric chemistry. They conclude that variations in transport alone cannot explain the recent slowing in the rate of decline in CFCs (Fig. 1), but that new emissions must have contributed.

Further evidence in support of their hypothesis comes from their observation of an increase in the difference between mean concentrations of atmospheric CFC-11 in the Northern and Southern hemispheres over the past few years — that is, the excess of CFC-11 in the Northern Hemisphere became larger. Moreover, after 2012, they observed the emergence of a strong relationship between the atmospheric concentration of CFC-11 and the concentrations of other ozone-depleting substances emitted as a result of human activities. These multiple lines of evidence support their conclusion that changes in atmospheric dynamics, especially in the stratosphere, must be acting in concert with renewed CFC-11 emissions to produce the observed concentrations. Such a careful analysis is crucial, because any claim of renewed — and therefore illegal — emissions will have political implications.

By taking into account the flow of the atmosphere to the locations at which the CFC-11 measurements were taken, Montzka *et al.* attribute the renewed emissions to east Asia. They also estimate that these emissions amount to about 13 gigagrams of CFC-11 per year (an increase of 25%) since 2012. However, the uncertainty in the inferred magnitude of fresh emissions might be up to 50%, mainly because of the difficulty in working out how air is transported between the stratosphere and the troposphere.

One way to reduce the uncertainties in the estimates and in the probable sources of the renewed emissions would be to use a high-resolution inverse-modelling approach, such as the one that has been used[8] in regional studies to attribute sources for emissions of hydrofluorocarbons (the chemicals that replaced CFCs and that are potential greenhouse gases). However, such an approach would probably need a denser global network of CFC measurements than is currently available. Moreover, regional inverse models would have to be extended to become global, high-resolution inverse models that include a well-resolved stratosphere and inter-hemispheric transport — which is a tall order, because few such comprehensive modelling systems are currently available that come close to the needed resolution.

Montzka and colleagues’ study highlights once more that environmental regulations cannot be taken for granted and must be safeguarded, and that monitoring is required to ensure compliance. Continuous observations of the environment are crucial: not only satellite measurements that yield global coverage, but also readings from measurement networks across the world that yield more-accurate *in situ* data. Taken together with models that encompass both the troposphere and the stratosphere, such data can be used to make defensible inferences about the sources of polluting chemicals. ■

Michaela I. Hegglin *is in the Department of Meteorology, University of Reading, Reading RG6 6BX, UK.*
e-mail: m.i.hegglin@reading.ac.uk

1. Montzka, S. A. *et al. Nature* **557**, 413–417 (2018).
2. Lovelock, J. E., Maggs, R. J. & Wade, R. J. *Nature* **241**, 194–196 (1973).
3. Molina, M. J. & Rowland, F. S. *Nature* **249**, 810–812 (1974).
4. Farman, J. C., Gardiner, B. G. & Shanklin, J. D. *Nature* **315**, 207–210 (1985).
5. World Meteorological Organization. *Scientific Assessment of Ozone Depletion: 2014* (WMO, 2014).
6. Annan, K. A. *We the Peoples: The Role of the United Nations in the Twenty-First Century* (UN, 2000).
7. Shepherd, T. G. *et al. Nature Geosci.* **7**, 443–449 (2014).
8. Brunner, D. *et al. Atmos. Chem. Phys.* **17**, 10651–10674 (2017).

STRUCTURAL BIOLOGY

Arresting vistas of arrestin activation

Computational and biochemical studies have revealed the mechanisms by which arrestin proteins are activated by G-protein-coupled receptors — potentially opening up broad avenues for drug discovery. SEE ARTICLE P.381 & LETTER P.452

BRIAN KRUMM & BRYAN L. ROTH

The largest family of drug targets in humans, and the principal therapeutic targets for at least 30% of approved medications in the United States, are the G-protein-coupled receptors[1] (GPCRs). When these transmembrane proteins detect extracellular agonist molecules, they transmit signals to the cell interior through G proteins inside the cells. The receptor is then sequentially phosphorylated to attenuate further signalling. The phosphorylated GPCR binds to an arrestin protein, and both undergo conformational changes that lead to the activation of arrestin-dependent cellular processes. Two papers in *Nature*, by Eichel *et al.*[2] and Latorraca *et al.*[3], now provide fresh insights into the mechanisms of arrestin activation and its consequences. Given the enormous potential of drugs that selectively target either G-protein or arrestin signalling, these findings might accelerate the development of safer and more-effective medications for a wide range of conditions.

Arrestins were first discovered in the visual system, where they bind to and inactivate a light-sensitive GPCR called rhodopsin[4]. They are now known to be almost universal regulators of GPCR signalling[5]. The binding of arrestin to GPCRs is enhanced by phosphorylation of the cytoplasmic tail — the carboxy terminus — of the receptors[6], and many models for arrestin binding and activation highlight the interaction of the protein with this region of the receptor.

It has been known since the 1990s that arrestins also bind at additional intracellular sites of several GPCRs, including the intracellular loops[7,8] (GPCRs have three intracellular loops that connect adjacent transmembrane regions of the receptor). In the past few years, structural[9] and biophysical studies[10] of arrestin bound to rhodopsin have clearly shown that arrestin binds to phosphorylated residues in the C terminus, as well as to a receptor core domain that includes intracellular loops 2 (IL2) and 3 (IL3). How these interactions lead to activation of arrestin and subsequent signalling has been obscure.

Latorraca and colleagues now cast light on this issue. The authors began by performing extensive computational simulations of the molecular dynamics of arrestin, both alone and during its interaction with various regions of rhodopsin. Their results indicate that ‘active’ arrestin fluctuates between active and inactive states, and that the receptor core domain and the phosphorylated tail can individually stabilize arrestin’s active state. Moreover, the active

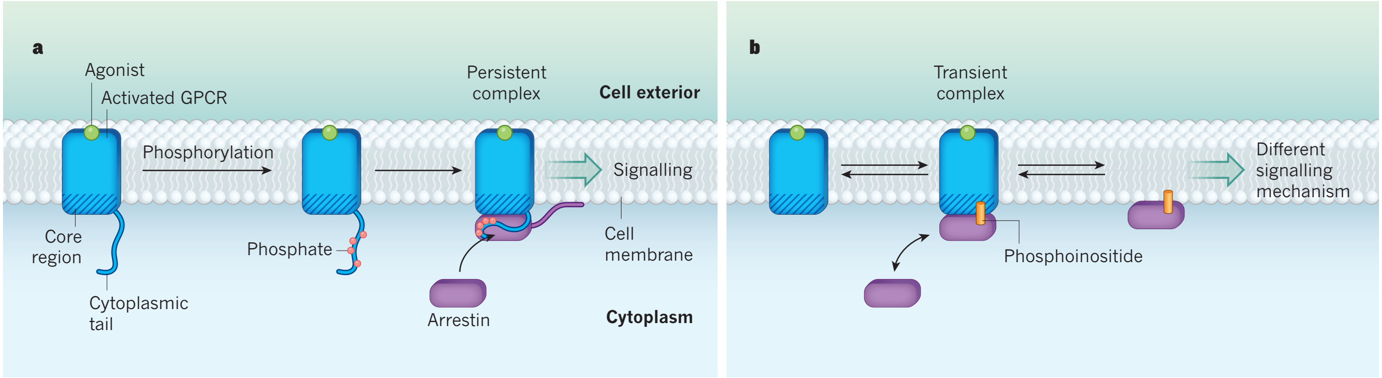


Figure 1 | Proposed mechanisms for arrestin activation. Arrestin proteins regulate the signalling of G-protein-coupled receptors (GPCRs) in cell membranes. Two papers[2,3] report computational and biochemical studies of GPCR–arrestin interactions, and propose mechanisms by which arrestin can be activated to redirect GPCR signalling. **a**, GPCRs are activated by the binding of agonist molecules. When the cytoplasmic tail of an activated GPCR is phosphorylated, arrestin is recruited to the receptor and interacts with both the cytoplasmic tail and the intracellular core region at the receptor's base. This activates arrestin, and leads to the formation of persistent GPCR–arrestin complexes that trigger arrestin-mediated signalling. A tail on arrestin that extends to proteins (not shown) in the cell membrane might also be needed for signalling. **b**, Alternatively, arrestin can interact transiently with the receptor core alone, in a process mediated by membrane-bound lipids known as phosphoinositides. The unbound, activated arrestin remains at the cell membrane, and goes on to trigger signalling through a different mechanism from that in **a**. (Figure adapted from Extended Data Fig. 9 in ref. 2.)

arrestin conformation is stabilized to an even greater extent when bound to both the core and the phosphorylated tail. The authors went on to perform further simulations of arrestin's interactions with the core domain. These suggest that a region of arrestin known as the finger loop stabilizes an interaction with the GPCR core domain, whereas interactions with IL2 and IL3 seem to trigger activation of arrestin.

A key insight from the molecular-dynamics simulations is that the two regions of arrestin to which the GPCR binds are allosterically coupled to each other: motions of the regions that bind the C terminus are coupled to motions of the regions that bind to the core, and vice versa. Importantly, the authors confirmed these computational findings directly in biophysical studies, using arrestin mutants tagged with fluorescent labels to monitor conformational changes at the protein surfaces that interact with the receptor's core and C terminus.

Latorraca and co-workers' simulations also suggest that the activated state of arrestin seems to persist even when the protein is not bound to the receptor. In their companion paper, Eichel *et al.* expand on and validate this prediction. The authors re-examined a phenomenon they had described previously[11], in which a GPCR known as the β1-adrenergic receptor (β1AR) interacts with arrestin only transiently when activated by an agonist. Arrestin then seems to be trafficked, independently of β1AR, to clathrin-coated endocytic structures (CCSs; vesicles that transport molecules into cells), where it can activate signalling proteins. This phenomenon is reminiscent of earlier findings that showed segregation of arrestins and GPCRs under some circumstances[12]. Such segregation was not anticipated by early models of GPCR–arrestin interactions, which posited that a stable GPCR–arrestin complex is essential for arrestin signalling[13].

Eichel and colleagues used a combination of microscopy techniques to show that transient engagement of the GPCR core, but not the C terminus, leads to prolonged accumulation of activated arrestin in CCSs. Such activation can be thought of as catalytic, because the GPCR activates arrestin but does not participate in subsequent downstream activation events — analogous to the way in which catalysts speed up reactions without directly taking part in them. The authors also demonstrated that the binding of phosphorylated lipids known as phosphoinositides at the cell membrane is essential for the capture of arrestin in CCSs after transient activation by GPCRs. The activated and GPCR-free arrestin can then activate downstream effectors such as ERKs.

The papers provide fresh insights that might accelerate the discovery and validation of biased GPCR agonists as therapeutic agents.

Taken together, these two papers open up a stunning new vista of GPCR–arrestin interactions in which an essential, bipartite interaction occurs through the receptor's core and phosphorylated C terminus (Fig. 1a). This engagement can lead to the formation of the persistent GPCR–arrestin 'scaffold' complexes that have frequently been observed and reported. An additional, transient form of this interaction leads to the catalytic activation of arrestin, which then accumulates in CCSs and triggers a specific form of arrestin signalling (Fig. 1b). It remains to be seen whether the mechanism of arrestin activation suggested by Latorraca and colleagues' molecular-dynamics simulations and by the structure of the arrestin–rhodopsin complex[9] is widely used by other arrestins and among GPCRs that do not undergo phosphorylation.

The idea of developing 'biased' GPCR agonists that selectively engage either G-protein or arrestin signalling[13] has tremendous potential for drug discovery. Such biased agonists have been proposed as safer and more-effective treatments for many disorders, including schizophrenia, chronic-pain conditions and heart disease[14,15]. By revealing cellular pathways for scaffold-based and catalytic arrestin activation, the current papers provide fresh insights that might accelerate the discovery and validation of biased GPCR agonists as therapeutic agents. ■

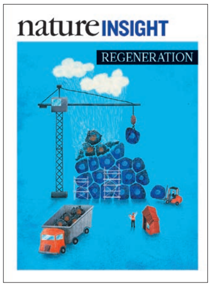
Brian Krumm *and* **Bryan L. Roth** *are in the Department of Pharmacology, School of Medicine, and the Division of Medicinal Chemistry and Chemical Biology, Eshelman School of Pharmacy, and in the NIMH Psychoactive Drug Screening Program, University of North Carolina at Chapel Hill, North Carolina 27705, USA.*
e-mail: bryan_roth@med.unc.edu

1. Sriram, K. & Insel, P. A. *Mol. Pharmacol.* **93**, 251–258 (2018).
2. Eichel, K. *et al. Nature* **557**, 381–386 (2018).
3. Latorraca, N. R. *et al. Nature* **557**, 452–456 (2018).
4. Wilden, U., Hall, S. W. & Kühn, H. *Proc. Natl Acad. Sci. USA* **83**, 1174–1178 (1986).
5. Luttrell, L. M. & Lefkowitz, R. J. *J. Cell Sci.* **115**, 455–465 (2002).
6. Palczewski, K., Buczyłko, J., Imami, N. R., McDowell, J. H. & Hargrave, P. A. *J. Biol. Chem.* **266**, 15334–15339 (1991).
7. Wu, G., Krupnick, J. G., Benovic, J. L. & Lanier, S. M. *J. Biol. Chem.* **272**, 17836–17842 (1997).
8. Gelber, E. I. *et al. J. Neurochem.* **72**, 2206–2214 (1999).
9. Kang, Y. *et al. Nature* **523**, 561–567 (2015).
10. Zhou, X. E. *et al. Cell* **170**, 457–469 (2017).
11. Eichel, K., Jullié, D. & von Zastrow, M. *Nature Cell Biol.* **18**, 303–310 (2016).
12. Bhatnagar, A. *et al. J. Biol. Chem.* **276**, 8269–8277 (2001).
13. Luttrell, L. M. *et al. Science* **283**, 655–661 (1999).
14. Wacker, D., Stevens, R. C. & Roth, B. L. *Cell* **170**, 414–427 (2017).
15. Smith, J. S., Lefkowitz, R. J. & Rajagopal, S. *Nature Rev. Drug Discov.* **17**, 243–260 (2018).

nature INSIGHT

REGENERATION

17 May 2018 / Vol 557 / Issue No 7705



Cover illustration
by Nik Spencer

The Campus
4 Crinan Street
London N1 9XW, UK
Tel: +44 (0) 20 7833 4000
e: nature@nature.com

SPRINGER NATURE

Restoration and repair are essential operations in the maintenance of the healthy human body. However, not all tissues heal at the same rate, if at all. This Insight on Regeneration examines the scope of cellular repair and the strategies currently being employed to challenge existing regenerative limitations.

In many cases of disease or injury it would be preferable to harness endogenous repair capacity. In the first Review, Wells and Watt discuss therapeutic options to stimulate cellular plasticity and drive regeneration while simultaneously avoiding malignant cellular transformation.

The brain possesses limited capacity for repair, with approaches such as cellular transplants producing only modest therapeutic results after injury. Barker, Götz and Parmer explore the use of stem-cell-derived structures and direct cellular reprogramming as powerful new tools for repairing neural circuits and function.

While these stem cell-based repair strategies hold much future potential, available options in the clinic are sparse. Blau and colleagues detail recent progress with biomaterials and platforms that aim to eliminate the current challenges of using stem cell-based treatments and to enhance their efficacy.

Despite decades of advances in the treatment of spinal cord injury, meaningful recovery remains elusive. In his Perspective, Michael Sofroniew explores the current state of spinal cord regeneration research and discusses why results and progress have been slow to come and controversial.

The two distinct components of the pancreas possess very different capacities for regeneration, with the exocrine pancreas exhibiting intrinsic repair, which the endocrine pancreas lacks. Zhou and Melton summarize this current knowledge and explore the therapeutic pipeline for combating pancreatic loss.

Treatments for restoring vision have been desired for centuries, but only recently has any significant progress been made. In the final Review, Roska and Sahel detail this recent progress and describe how the latest model systems and translational strategies will generate new opportunities to combat vision loss.

We hope this diverse collection stimulates further discussion and research in the field of regenerative medicine.

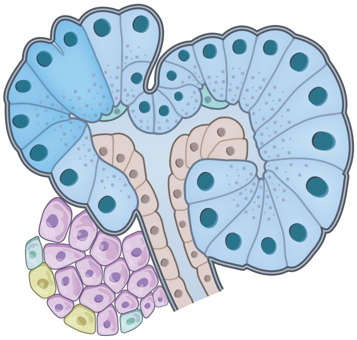
Noah Gray, Nathalie Le Bot & Marie-Thérèse Heemels
Senior Editors

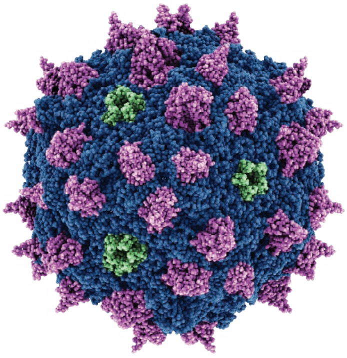
CONTENTS

REVIEWS

PERSPECTIVE

REVIEWS





REVIEW

https://doi.org/10.1038/s41586-018-0073-7

Diverse mechanisms for endogenous regeneration and repair in mammalian organs

James M. Wells[1]* & Fiona M. Watt[2]*

Mammalian organs comprise an extraordinary diversity of cell and tissue types. Regenerative organs, such as the skin and gastrointestinal tract, use resident stem cells to maintain tissue function. Organs with a lower cellular turnover, such as the liver and lungs, mostly rely on proliferation of committed progenitor cells. In many organs, injury reveals the plasticity of both resident stem cells and differentiated cells. The ability of resident cells to maintain and repair organs diminishes with age, whereas, paradoxically, the risk of cancer increases. New therapeutic approaches aim to harness cell plasticity for tissue repair and regeneration while avoiding the risk of malignant transformation of cells.

The field of cell and gene therapy is now starting to mature, as the number of reports of clinical safety and efficacy in patients has begun to increase[1]. Although the number of patients treated is small, the use of autologous human induced pluripotent stem cells as a treatment for macular degeneration appears to be safe[2], and the ability of gene-corrected autologous epidermal cells to provide a long-lasting cure for a skin-blistering disease has been established[3]. The commercial and academic sectors have also been buoyed by the successes of cancer immunotherapies such as CAR T cell treatments[4].

Cell, tissue and organ replacement is needed to treat irreversible organ failure. However, in many cases of damage or disease, a better strategy might be to maintain and reinvigorate organ function in situ. This is an attractive option because it is likely to be less invasive and more cost-effective than transplantation. Stimulating endogenous repair by manipulating stem cells and/or their niche is of growing interest, particularly as we come to understand more about the components of the cellular microenvironment[5] and start to exploit the recently revealed plasticity of differentiated cells in adult tissues[6].

Here we discuss different strategies for tissue repair and regeneration and consider the mechanisms by which cell plasticity is induced in adult tissues. Next, we consider the extent to which induced plasticity might be beneficial in reversing tissue ageing and the potential risk of plasticity-associated cancer. Finally, we remark on the feasibility of stimulating endogenous repair and regeneration to treat disease.

Some definitions

The stem cell field has evolved to the point at which past terminology and dogma are outdated in the context of the adult mammal and so it is worth revisiting some of the terminology. 'Repair' is probably the most straightforward term because it can be defined as restoring damaged tissue to good condition following insults such as ageing, disease and injury. Tissues differ widely in their ability to repair; the skin and gastrointestinal tract, for example, repair very effectively, whereas the brain is a tissue in which repair is highly inefficient. In the case of skin (Fig. 1), although repair efficiently restores the barrier properties of the tissue, the repaired tissue may not be entirely normal; there may, for example, be scar formation and the repaired skin may be devoid of hair follicles. In situations in which new hair follicles form following wounding[7,8], we consider them to have regenerated (Fig. 1). We consider the continuous replacement of epithelial cells that are shed from the surface of the skin or the lumen of the gut to constitute 'regeneration', essentially a high rate of cell turnover under steady-state conditions. Therefore, regeneration is not obligatorily linked to tissue repair.

We now appreciate that daughter cells in several organs can exit the stem cell compartment and become increasingly lineage-restricted yet still retain the potential to revert to a stem cell state. This ability of a more differentiated cell to give rise to a stem cell, or the interconversion of distinct stem cell populations, constitutes 'plasticity' and has strong parallels with the terms 'dedifferentiation' and 'transdifferentiation' that are well known from classical embryology[9] (Fig. 2). Furthermore, the phenomenon of plasticity suggests that the historical concepts of cell fate specification and determination are not, as previously thought, irreversible. Plasticity also calls into question the meaning of 'terminal differentiation', which is classically used to describe a unidirectional process by which post-mitotic differentiated cells are formed from a stem cell population. When such a cell lacks a nucleus, as is the case for human red blood cells or cells in the outer layers of human epidermis, there is no doubt that the state of differentiation is 'terminal', meaning 'irreversible'. However, there are now many examples across organ systems in which cells that are committed to a specific lineage can re-enter the stem cell compartment after injury[6,10]. Therefore, it would appear that even if the trajectory of a cell is towards a terminal state, different steps along the route may be reversible.

Finally, what is meant by 'quiescence'? Historically, this refers to cells that are not actively dividing and may have arrested in the G0 phase of the cell cycle. If such cells incorporate a DNA label in S phase, they may retain the label over many months or even years; however, these DNA label-retaining cells are not obligatorily arrested in G0[11]. In stem cell biology, quiescence is now generally taken to mean a dormant genome (heterochromatinization and low levels of global transcription) and metabolic dormancy (low protein translation and lack of oxidative phosphorylation)[12,13].

The caveats of cell lineage tracing

Inevitably, much of our understanding of the nature and properties of stem cells in adult tissues comes from lineage-tracing studies in mice.

---

[1]Divisions of Developmental Biology and Endocrinology, Center for Stem Cell and Organoid Medicine, Cincinnati Children's Hospital Medical Center, Cincinnati, Ohio, USA. [2]Centre for Stem Cells and Regenerative Medicine, Guy's Hospital Campus, King's College London, London, UK. *e-mail: james.wells@cchmc.org; fiona.watt@kcl.ac.uk

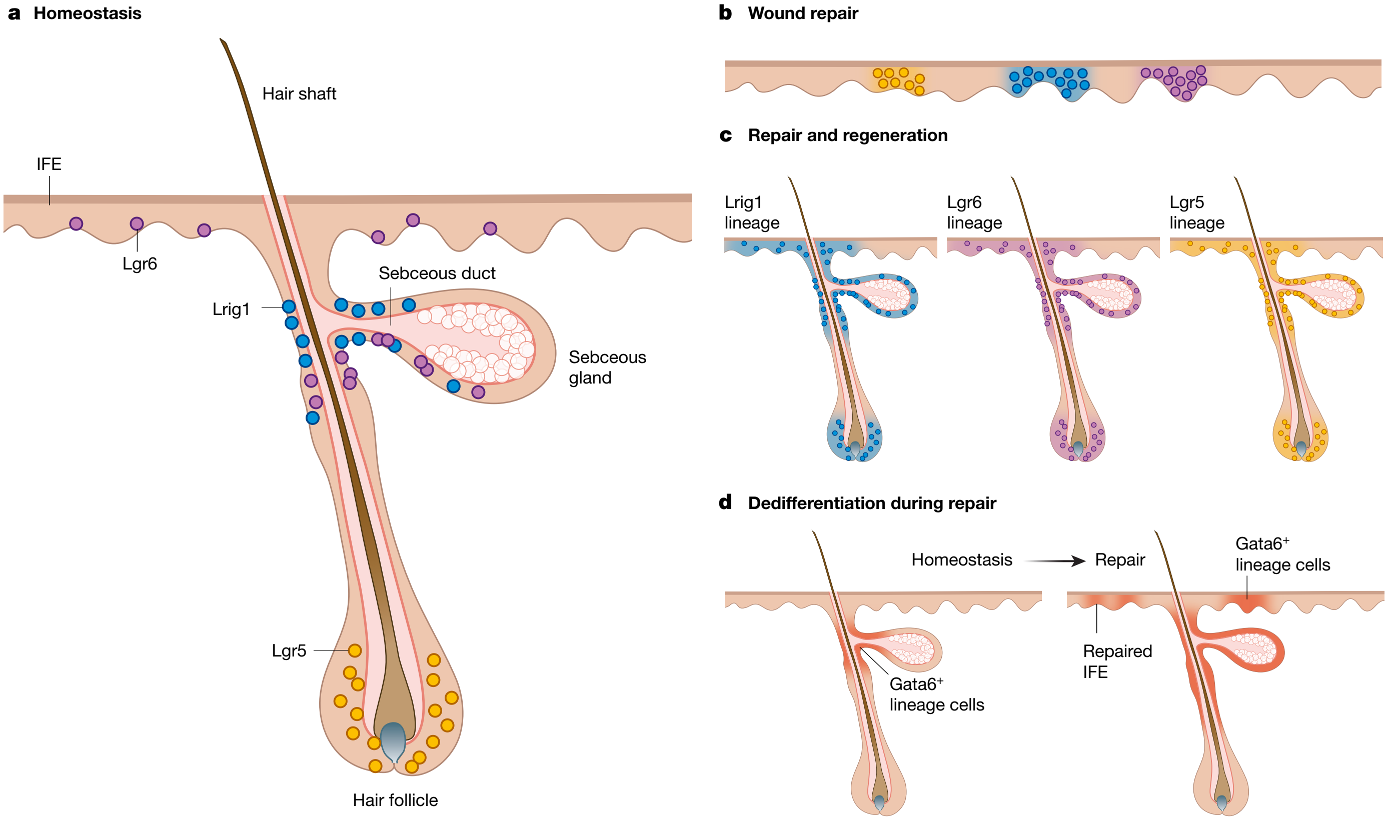


Fig. 1 | Epidermal plasticity. a, Location of Lgr5$^+$, Lgr6$^+$ and Lrig1$^+$ stem cells in undamaged adult mouse dorsal epidermis. **b**, **c**, After a wound occurs, the interfollicular epidermis (IFE) may be regenerated in the absence of hair follicles and sebaceous glands (**b**, epidermal repair) or the interfollicular epidermis and adnexal structures may re-form (**c**, epidermal regeneration). Lgr5, Lgr6 and Lrig1 lineages can contribute to both repair and regeneration. **d**, After a wound occurs, differentiated cells of the Gata6 lineage that are normally confined to the suprabasal layers of the sebaceous duct are able to enter the wound site, undergo dedifferentiation and contribute to the stem cell compartment of the interfollicular epidermis and to the hair follicle and sebaceous gland.

This is a powerful technology, which enables us not only to trace the progeny of individual cells over many generations, but also to selectively ablate or expand specific populations. However, there are some technical constraints that must not be ignored. One is so elementary that it should not need to be discussed: it is unwise to assume that a marker that is expressed in a particular stem cell population is not also expressed in other cells elsewhere within the tissue. A good example of this is Lrig1, which is an excellent epithelial stem cell marker in a range of tissues, including the epidermis; however, it is also expressed in a highly dynamic manner in a subset of skin fibroblasts[14]. Except in pathological situations, in which epithelial cells can potentially invade the underlying connective tissue or undergo epithelial–mesenchymal transition, the gene of interest can still provide a highly useful reporter. Indeed, Lgr5 and Lgr6, which are often used as stem cell markers in a range of epithelia, are also expressed by mesenchymal populations in the lung[15].

Co-expression of the stem cell marker Lgr5 in differentiated cells or de novo expression following injury may have led to some confusion as to the source of regenerating cells. Following injury, increased numbers of Lgr5-lineage-traced cells have been seen in proliferative organs, such as the stomach[16] and skin, as well as relatively slow cycling organs, such as the liver[17] and pancreas[18]. In the stomach, Lgr5 is expressed in stem cells as well as chief cells following injury, and this highlights an ongoing debate on whether chief cells contribute to the regenerating gastric epithelium (Box 1).

The use of dual recombinase systems can overcome this weakness since lineage tracing relies on co-expression of two factors in the same cell[19]. This approach was used to revisit the concept that biliary epithelial cells can transdifferentiate into hepatocytes following injury, a concept for which there are contradictory results (Box 1). By using two separately encoded recombinase proteins (Sox9-CreER and Alb-DreER), researchers have demonstrated that Sox9-expressing biliary epithelial cells did not form hepatocytes after injury. This is consistent with the emerging concept that biliary epithelial cells give rise to hepatocytes only under conditions in which hepatocytes are unable to divide[20,21].

Other approaches for lineage tracing include barcoding and exploitation of naturally occurring DNA mutations. For example, an artificial DNA recombination locus (called Polylox) was recently developed for barcoding based on Cre–*loxP* recombination[22]. Polylox recombination in situ generates several hundred thousand barcodes, enabling the tagging of single cells in mice. These barcodes were then used for fate mapping of haematopoietic stem cells, which demonstrated that the adult haematopoietic stem cell compartment is a mosaic of clones that were present in the embryo and that most clones in the adult have multilineage differentiation potential. Newly developed CRISPR–Cas-based systems can be used to generate large-scale maps of cell lineages in multicellular systems for analyses of normal development and disease[23].

In terms of human lineage tracing, sun-exposed human skin has been a tractable system for many years and recent studies have inferred the clonal architecture of the epidermis based on DNA sequencing[24,25]. In this case, a caveat that has emerged is that the interpretation of the data depends on how much skin is sequenced to capture the largest mutant clones. When sufficiently large areas of skin are sequenced, it is possible to infer the effects of primary and secondary mutations, leading to the conclusion that secondary mutations arising at the edge of a mutant clone have a selective growth advantage[26]. Large-scale single-cell gene expression profiling of adult human tissues will

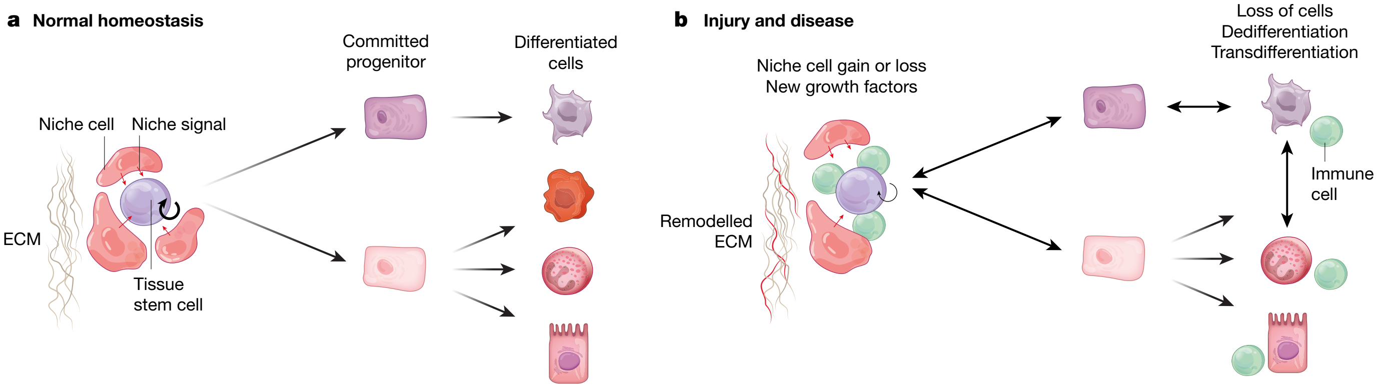


Fig. 2 | General mechanisms of plasticity. **a**, Under normal homeostatic conditions stem cells self-renew and generate differentiated lineages in a unidirectional manner via committed progenitors. The fate of stem cells—self-renewal or generation of differentiated cells - is influenced by signals from the niche, including signals from specific cell types, soluble signals and extracellular matrix (ECM). **b**, In the context of tissue injury or disease, stem and progenitor cells and their differentiated progeny exhibit plasticity, including dedifferentiation and transdifferentiation. This is driven, in part, by altered niche signals, including new interactions with cells of the immune system and ECM remodelling.

undoubtedly improve our understanding of the inter-relationships between cell states, cell types and lineages[27].

Cell-intrinsic mechanisms of plasticity

Plasticity of adult tissues can be apparent as dedifferentiation or transdifferentiation from one cell type into another[28] (Fig. 2). Although in most cases plasticity is linked to injury, this is not an obligatory association. In the epidermis, for example, switching between stem cell populations and fate switching between differentiated lineages can be induced by β-catenin activation in the undamaged epidermis[29]. It is possible that this occurs because the epidermis in which β-catenin is activated can reprogram the underlying connective tissue, the dermis, to a neonatal state by secretion of signalling molecules such as sonic hedgehog[8,30]. In support of this, the hair follicle defects that result from deleting the transcriptional co-activator Blimp1 in the dermis are rescued by epidermal β-catenin activation[31].

There are also numerous examples of the use of transcription factors to induce plasticity by cell-autonomously modulating gene expression. This is evident in the pancreas, which exhibits very limited regenerative potential under homeostatic conditions, yet harbours a latent multipotent cell type that can be activated[32]. Expression of endocrine-promoting transcription factors in the exocrine cells of the pancreas triggers their conversion to insulin-expressing cells capable of reversing diabetes[33]. Inactivation of the SCF-type E3 ubiquitin ligase substrate recognition component Fbw7 induces pancreatic ductal cells to reprogram into α, δ and β cells and acts by stabilizing the transcription factor Neurog3, a known regulator of endocrine cell differentiation[34].

Injury-induced plasticity

The induction of pluripotency by the expression of transcription factors demonstrates that plasticity can be experimentally induced in almost all somatic cells[98]. However, in nature, plasticity is most commonly associated with tissue injury and damage (Box 1). Injuries can occur in a number of ways, such as by targeted laser ablation[35], tissue disaggregation[10] or full-thickness wounds[7,36] in the skin. In the intestine, diphtheria toxin receptor (DTR)-mediated stem cell ablation[37] or 5-fluoro-uracil-mediated killing of proliferative cells[38] causes secretory and absorptive progenitors to acquire intestinal stem cell-like properties and contribute to regeneration. A rare population of cells in the crypt that appears to consist of enteroendocrine cells also expresses stem cell markers and can broadly contribute to epithelial regeneration following irradiation[39]. Although some experimental forms of injury such as DTR-mediated killing of genetically marked cells are not normal physiological conditions, the effects of *Helicobacter pylori* in the stomach are. Chronic injury caused by *H. pylori* infection can directly amplify the regenerative response and cause a number of epithelial changes, including loss of parietal cells and confusion of cellular identity of chief cells[40].

Plasticity is not limited to the epithelial compartment of tissues. For example, regeneration of adipocytes from dermal myofibroblasts during wound healing in the skin has recently been demonstrated[41]. This occurs through the interaction with hair follicles that are forming in the wound. The hair follicles trigger bone morphogenetic protein (BMP) signalling and activation of adipocyte transcription factors. Indeed, it has been suggested that fibrosis occurs not only as a consequence of aberrant signalling but also because of aberrant differentiation of a particular subpopulation of fibroblasts[42].

Inflammation and infection-induced plasticity

Microbe- and injury-induced inflammatory responses are mediated by cells of the innate immune system, which produce and respond to cytokine mediators and other inflammatory signals. Tissue-resident cells sense these signals and respond by migrating, proliferating and regenerating the tissue[43]. Mechanisms linking inflammation to damage repair and regeneration in mammals are conserved in lower organisms[44]. In terms of immune modulation of stem cells and regeneration, evolutionary advances in the immune system appear to be inversely correlated with the ability to fully regenerate injured tissues[45].

There are a number of ways in which the immune system can facilitate endogenous repair. For example, there is a strong link between macrophage-mediated debris clearance and regeneration in the central nervous system[45] and macrophages are essential for effective wound healing in the skin[46]. In addition, immune cells can activate proliferation: in undamaged skin, Notch-mediated communication between regulatory T cells and hair follicle stem cells is required for hair follicle growth[47]. Regulatory T cells can also activate muscle satellite cells following injury[45]. In addition, TIE2-expressing monocytes/macrophages promote revascularization of ischaemic limbs[48].

Inflammation also facilitates repair through the modulation of stem cell regulatory proteins such as Lgr5. Lgr5 is a Wnt target gene and macrophages are a known source of Wnt ligands in a variety of contexts including ulcerative colitis[49]. Lgr5 is also directly regulated by inflammatory mediators. In basal cell carcinoma, IKKα directly binds to the Lgr5 promoter and regulates transcription of Lgr5[50]. In the intestine, interleukin-22 acts via Stat3 to promote epithelial regeneration[51]. Therefore, immune cells could mediate an increase in plasticity through direct regulation of stem cell pathways.

In addition to these roles for inflammation in tissue repair, recent data suggest that inflammation can trigger epigenetic memory of injury[52] and that innate immune signalling pathways can have a role in nuclear reprogramming (transflammation[53]), for example via reactive oxygen species. If inflamed skin is wounded, heals and is then re-wounded, the second wound heals more rapidly[52]. This is because epidermal stem cells maintain chromosomal accessibility to key

BOX 1

Examples of plasticity

Lung. There are two major compartments in the adult lung: the airways and the alveoli[28]. In the airway epithelia, differentiated club cells (secretory cells) can directly convert into ciliated cells. Basal cells replenish all types of secretory and ciliated cells after tissue damage. Club cells can dedifferentiate into basal stem cells after stem cell ablation. The alveolar epithelium consists of type-1 cells that permit gas exchange and surfactant-producing type-2 cells. Type-2 cells generate type-1 and type-2 cells under homeostatic conditions and after injury. In addition, after injury type-1 cells can generate type-2 stem cells.

Stomach. The stomach is regionalized into the proximal corpus and distal antrum (pylorus). The corpus epithelium contains differentiated chief and parietal cells, whereas the antrum is distinguished by gastrin-producing G cells. Antrum stem cells express Lgr5[88,89], whereas corpus stem cells express Sox2, Lrig1 and Mist1[90–92]. Chief cells are long-lived, highly specialized, pepsinogen-producing cells and become recruited to contribute to epithelial repair following injury[16,38].

Liver. The liver largely consists of two cell types, hepatocytes and cholangiocytes. A subpopulation of Wnt-responsive hepatocytes maintains the liver under steady-state conditions[93]. After a partial hepatectomy, the remaining hepatocytes divide until the original liver mass is restored and then return to a non-proliferative state[94]. By contrast, when there is acute or chronic damage to hepatocytes such that they are unable to proliferate, cholangiocyte-like cells near the bile duct tree regenerate hepatocytes and cholangiocytes[20,21].

Similarities in regeneration and plasticity across gastrointestinal organs. There are vast structural and functional differences in the epithelium of gastrointestinal organs. The stratified squamous epithelium of the oesophagus has little cellular diversity, whereas the stomach is a complex glandular mucosa, and the intestine consists of a diverse mix of absorptive and secretory cell types. Nevertheless, these organs share many features. The entire epithelium of the gastrointestinal tract derives from a single layer of embryonic endoderm[95] and similar signalling pathways control epithelial homeostasis in the different organs. Canonical Wnt signalling is involved in maintaining and expanding stem and progenitor cells, whereas BMP and Notch signalling promote differentiation of the various epithelial lineages. Indeed, there is evidence that the caudal homeobox transcription factor Cdx2 is the single factor that functionally distinguishes an intestinal from a gastric stem cell[96,97]. It is not clear why gastrointestinal stem cells have so many similarities, but it could be a primitive epigenetic memory owing to their shared cellular origin during embryonic development.

stress response genes that are activated by the original inflammatory stimulus. This memory depends on the intracellular DNA sensor absent in melanoma 2 (*Aim2*), which encodes an activator of the inflammasome. In the absence of Aim2 or its downstream effectors the epidermal memory is erased. More broadly, intracellular nucleic acid sensors can discriminate between self and non-self nucleic acids[54], opening up the possibility that pathogens can induce plasticity in adult tissues by a similar mechanism.

Cell–extracellular matrix and cell–cell adhesion

In addition to inducing inflammation, injury induces changes in the adhesive interactions between cells and their environment. Epithelial damage can lead to loss of basement membrane proteins and deposition of new extracellular matrix (ECM) with a different protein composition. It can also lead to loss of neighbouring cells and interactions with new neighbours, linked, for example, to cell migration[10]. Inevitably, the fine patterning of niches, leading to local and spatially distinct signals in undamaged tissue[55], is destroyed after tissue damage.

Recent studies have demonstrated that physical factors play a key part in determining tissue architecture under steady-state conditions. For example, the undulations of the human epidermal–dermal junction are instructive in epidermal stem cell clustering[56]. Crowding in the epidermal basal layer affects cell shape; a decrease in cortical tension, and increased cell–cell adhesion trigger differentiation and movement of cells into the suprabasal layers[57]. Piezo1 has a role in linking mechanical stretch to division of epithelial cells[58]. One of the key integrators of cell position and cell fate in the epidermis and other tissues is YAP activity[59–61]. YAP integrates diverse signals, including Notch, Wnt and ECM signals[61,62], and intercellular adhesion[60]. In mice, YAP has a key role in cell fate transitions during colonic regeneration following injury[62]. The colonic epithelium is transiently reprogrammed into a primitive, fetal-like state via ECM remodelling, FAK–Src signalling and YAP activation.

One general mechanism by which injury can induce plasticity is by modulation of the actin cytoskeleton. For example, Rho-kinase inhibition expands epithelial stem cells in culture, including cells from the mammary gland, prostate, intestine and colon[63,64]. Similarly, inhibition of dual SMAD signalling enables long-term expansion of basal cells in a variety of epithelia, both in vivo and in vitro[65]. TGFβ–BMP–SMAD signalling is activated in differentiated cells of multiple epithelia, including lung, mammary gland, stomach and epidermis, and is suppressed in basal cells that express the transcription factor p63. In lung airway epithelium, SMAD signalling promotes differentiation, whereas SMAD inhibition leads to stem cell hyperplasia. These findings are likely to be relevant in epithelial cancers. For example, suprabasal expression of the $\alpha6\beta4$ integrin in mouse epidermis increases tumour formation in part by relieving TGFβ-mediated growth inhibition[66].

Ageing and regeneration

Tissue integrity, organ function and regenerative capacity all decline with age[67]. Despite an obvious relevance to human health, there are relatively few studies about changes in regeneration/homeostasis with age, possibly due to the technical and financial burdens of studying ‘old’ animals. Factors that are likely to be involved include loss of stem cell number and proliferative capacity, ECM remodelling towards a more fibrotic phenotype, and cumulative DNA damage causing cellular senescence (Fig. 3). It is clear that stem cell extrinsic factors, such as fibrosis caused by chronic inflammation, accelerate these processes. However there are also age-linked cell-intrinsic changes in stem and progenitor cells. Age-related changes in tissue-resident stem cells have been observed in a range of tissues, including blood, skin and intestine[67].

Some age-related changes are cell-intrinsic, since stem cells isolated from aged patients grow poorly in culture relative to their young counterparts. The cell-intrinsic mechanisms that drive stem cell ageing include loss of telomeres, epigenetic changes[68], decline in mitochondrial function[69] and accumulation of DNA mutations[70]. In the intestine, telomere shortening is linked to Wnt inactivity and stem cell failure[71]. The cellular safeguards to cancer, including programmed cell death, senescence and differentiation, might be responsible for reduced stem cell number and activity as we age. How mutations are coupled to the number of stem cell replications has been discussed previously[70] as well as how this might cause the well-known increase in cancer risk with age. An increase in mutation load in stem cells would be expected to correlate with increased cell death and senescence, resulting in the loss of stem cell number or proliferative capacity.

The therapeutic possibility of rejuvenating aged organs is best exemplified by parabiosis experiments. Parabiosis, the surgical linking of the circulatory systems between old and young animals, results in the restoration of stem cell activity in the organs of old mice. Moreover, a single transfusion of old blood causes a rapid loss of stem cell activity in young mice, demonstrating the negative effects of circulating factors in old animals[72,73]. The effects of parabiosis might directly act on stem

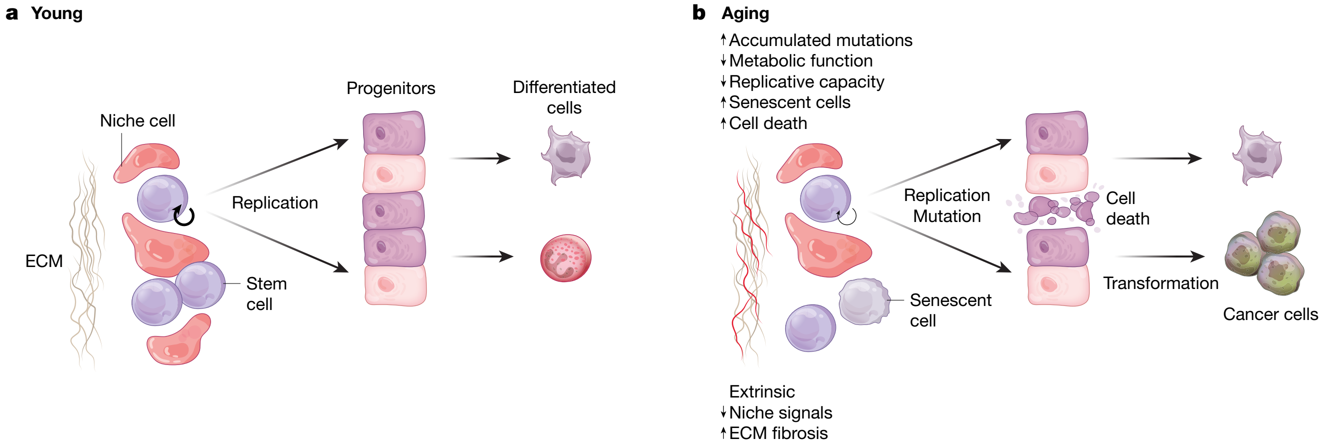


Fig. 3 | General mechanisms of ageing. The balance between stem cell renewal and differentiation and the interactions between stem cells and the niche (**a**) are altered with ageing (**b**). Age-associated changes are both cell-intrinsic, such as the accumulation of mutations, and niche-associated, such as fibrotic ECM.

cells[74] or might act to remodel and rejuvenate the niche[75]. Circulating factors that might modulate the niche include ECM proteins, growth factors and immune cells. In the skin, age-associated proteolysis of type-XVII collagen results in hair follicle miniturization and epidermal hypertrophy, which can be reversed by re-expressing that collagen subtype[76,77]. Soluble growth factors and cytokines, such as Wnt and interleukin-22, can reverse age-related declines in intestinal stem cells[51,78]. One of the most recent approaches to rejuvenate organs is through the targeted removal of senescent cells in the niche[79].

Plasticity and cancer

The obvious downside of inducing plasticity for tissue rejuvenation is the risk of cancer. Inflammation-induced lineage infidelity in the epidermis is transient in wounds but persists in cancer[36]. Specific chromatin states prime cells in squamous cell carcinoma to undergo epithelial–mesenchymal transition[80]. A further finding is that neighbouring wild-type cells have the ability to correct aberrant epidermal behaviour of cells that have β-catenin mutations[81].

There are multiple examples linking cellular plasticity and cancer in the gastrointestinal tract. In the stomach, induction of plasticity in differentiated cells occurs during chronic injury, in which chief cells co-express mucous neck cell markers (TTF2, which is also known as spasmolytic polypeptide) in a precancerous lesion termed SPEM (spasmolytic polypeptide expressing metaplasia)[82]. A further example of cancer-associated plasticity is intestinal metaplasia, in which the oesophagus or stomach epithelium converts to an intestinal epithelium. In the case of Barrett's oesophagus, an intestine-like metaplasia and precursor of oesophageal adenocarcinoma is triggered by gastro-oesophageal reflux. Recently it has been suggested that transitional basal cells at the squamous–columnar junction generate Barrett's oesophagus[83].

Therapeutic prospects

Treatments that target the niche are already being evaluated to treat cancer[5,84] and this Review has highlighted a number of strategies that could be exploited to promote tissue regeneration. Looking into the future, computational analysis of cell state transitions indicates that it may be possible to pharmacologically stabilize particular cell states—whether stem cell, committed cell or differentiated cell—for therapeutic benefits[85]. In addition, mining large datasets of single-cell gene expression profiles[86] is likely to enhance our understanding of different cell types and states and the pathways by which they are specified.

One concern is whether the induction of plasticity will ever be therapeutically relevant, given that the proportion of cells in a tissue that exhibit plasticity is often exceptionally rare[37,38]. Time will tell, but it is worth concluding with the observation that gastric bypass surgery in humans enhances pancreatic β-cell function and may also increase β-cell mass[87]. While it is not clear whether this involves injury-induced plasticity, it is proof that a surgical intervention can have unanticipated therapeutic benefits.

Received: 22 January 2018; Accepted: 7 March 2018;
Published online 16 May 2018.

1. Cossu, G. et al. *Lancet* commission: stem cells and regenerative medicine. *Lancet* **391**, 883–910 (2017).
2. Mandai, M. et al. Autologous induced stem-cell-derived retinal cells for macular degeneration. *N. Engl. J. Med.* **376**, 1038–1046 (2017).
3. Hirsch, T. et al. Regeneration of the entire human epidermis using transgenic stem cells. *Nature* **551**, 327–332 (2017).
A combination of gene and cell therapy is used to repair most of the epidermis of a child with a skin blistering disorder.
4. Lim, W. A. & June, C. H. The principles of engineering immune cells to treat cancer. *Cell* **168**, 724–740 (2017).
5. Lane, S. W., Williams, D. A. & Watt, F. M. Modulating the stem cell niche for tissue regeneration. *Nat. Biotechnol.* **32**, 795–803 (2014).
6. Rajagopal, J. & Stanger, B. Z. Plasticity in the adult: how should the Waddington diagram be applied to regenerating tissues? *Dev. Cell* **36**, 133–137 (2016).
7. Ito, M. et al. Wnt-dependent de novo hair follicle regeneration in adult mouse skin after wounding. *Nature* **447**, 316–320 (2007).
8. Rognoni, E. et al. Inhibition of β-catenin signalling in dermal fibroblasts enhances hair follicle regeneration during wound healing. *Development* **143**, 2522–2535 (2016).
9. Merrell, A. J. & Stanger, B. Z. Adult cell plasticity in vivo: de-differentiation and transdifferentiation are back in style. *Nat. Rev. Mol. Cell Biol.* **17**, 413–425 (2016).
10. Donati, G. et al. Wounding induces dedifferentiation of epidermal Gata6$^+$ cells and acquisition of stem cell properties. *Nat. Cell Biol.* **19**, 603–613 (2017).
Demonstration that terminally differentiated cells of the sebaceous duct can dedifferentiate and contribute to long-term repopulation of the interfollicular epidermis during wound repair.
11. Li, N., Nakauka-Ddamba, A., Tobias, J., Jensen, S. T. & Lengner, C. J. Mouse label-retaining cells are molecularly and functionally distinct from reserve intestinal stem cells. *Gastroenterology* **151**, 298–310 (2016).
12. Rodgers, J. T. et al. mTORC1 controls the adaptive transition of quiescent stem cells from G_0 to G_{Alert}. *Nature* **510**, 393–396 (2014).
13. Signer, R. A., Magee, J. A., Salic, A. & Morrison, S. J. Haematopoietic stem cells require a highly regulated protein synthesis rate. *Nature* **509**, 49–54 (2014).
14. Driskell, R. R. et al. Distinct fibroblast lineages determine dermal architecture in skin development and repair. *Nature* **504**, 277–281 (2013).
15. Lee, J. H. et al. Anatomically and functionally distinct lung mesenchymal populations marked by Lgr5 and Lgr6. *Cell* **170**, 1149–1163 (2017).
16. Leushacke, M. et al. Lgr5-expressing chief cells drive epithelial regeneration and cancer in the oxyntic stomach. *Nat. Cell Biol.* **19**, 774–786 (2017).
17. Cao, W. et al. Dynamics of proliferative and quiescent stem cells in liver homeostasis and injury. *Gastroenterology* **153**, 1133–1147 (2017).
18. Huch, M. et al. Unlimited in vitro expansion of adult bi-potent pancreas progenitors through the Lgr5/R-spondin axis. *EMBO J.* **32**, 2708–2721 (2013).
19. He, L. et al. Enhancing the precision of genetic lineage tracing using dual recombinases. *Nat. Med.* **23**, 1488–1498 (2017).
A lineage tracing strategy that depends on the co-expression of two different recombinases can target more specific populations of stem and progenitor cells.
20. Lu, W. Y. et al. Hepatic progenitor cells of biliary origin with liver repopulation capacity. *Nat. Cell Biol.* **17**, 971–983 (2015).

21. Raven, A. et al. Cholangiocytes act as facultative liver stem cells during impaired hepatocyte regeneration. *Nature* **547**, 350–354 (2017).
When hepatocyte proliferation is impaired, biliary cells can contribute to regeneration of both bile ducts and hepatocytes.
22. Pei, W. et al. Polylox barcoding reveals haematopoietic stem cell fates realized in vivo. *Nature* **548**, 456–460 (2017).
The authors use a Cre–*loxP* approach to genetically barcode thousands of individual haematopoietic stem cells and follow their descendants in vivo.
23. McKenna, A. et al. Whole-organism lineage tracing by combinatorial and cumulative genome editing. *Science* **353**, aaf7907 (2016).
24. Martincorena, I. et al. High burden and pervasive positive selection of somatic mutations in normal human skin. *Science* **348**, 880–886 (2015).
25. Simons, B. D. Deep sequencing as a probe of normal stem cell fate and preneoplasia in human epidermis. *Proc. Natl Acad. Sci. USA* **113**, 128–133 (2016).
26. Lynch, M. D. et al. Spatial constraints govern competition of mutant clones in human epidermis. *Nat. Commun.* **8**, 1119 (2017).
The size of some clones in cancer-prone epidermis is too large to be accounted for by neutral drift.
27. Regev, A. et al. The human cell atlas. *eLife* **6**, e27041 (2017).
28. Tata, P. R. & Rajagopal, J. Plasticity in the lung: making and breaking cell identity. *Development* **144**, 755–766 (2017).
29. Kretzschmar, K., Weber, C., Driskell, R. R., Calonje, E. & Watt, F. M. Compartmentalized epidermal activation of β-catenin differentially affects lineage reprogramming and underlies tumor heterogeneity. *Cell Rep.* **14**, 269–281 (2016).
30. Lichtenberger, B. M., Mastrogiannaki, M. & Watt, F. M. Epidermal β-catenin activation remodels the dermis via paracrine signalling to distinct fibroblast lineages. *Nat. Commun.* **7**, 10537 (2016).
31. Telerman, S. B. et al. Dermal Blimp1 acts downstream of epidermal TGFβ and Wnt/β-catenin to regulate hair follicle formation and growth. *J. Invest. Dermatol.* **137**, 2270–2281 (2017).
32. Demcollari, T. I., Cujba, A. M. & Sancho, R. Phenotypic plasticity in the pancreas: new triggers, new players. *Curr. Opin. Cell Biol.* **49**, 38–46 (2017).
33. Zhou, Q. et al. A multipotent progenitor domain guides pancreatic organogenesis. *Dev. Cell* **13**, 103–114 (2007).
34. Sancho, R., Gruber, R., Gu, G. & Behrens, A. Loss of Fbw7 reprograms adult pancreatic ductal cells into α, δ, and β cells. *Cell Stem Cell* **15**, 139–153 (2014).
35. Rompolas, P., Mesa, K. R. & Greco, V. Spatial organization within a niche as a determinant of stem-cell fate. *Nature* **502**, 513–518 (2013).
36. Ge, Y. et al. Stem cell lineage infidelity drives wound repair and cancer. *Cell* **169**, 636–650 (2017).
37. Tetteh, P. W. et al. Replacement of lost *Lgr5*-positive stem cells through plasticity of their enterocyte-lineage daughters. *Cell Stem Cell* **18**, 203–213 (2016).
38. Stange, D. E. et al. Differentiated *Troy*$^{+}$ chief cells act as reserve stem cells to generate all lineages of the stomach epithelium. *Cell* **155**, 357–368 (2013).
39. Yan, K. S. et al. Intestinal enteroendocrine lineage cells possess homeostatic and injury-inducible stem cell activity. *Cell Stem Cell* **21**, 78–90 (2017).
40. Mills, J. C. & Sansom, O. J. Reserve stem cells: differentiated cells reprogram to fuel repair, metaplasia, and neoplasia in the adult gastrointestinal tract. *Sci. Signal.* **8**, re8 (2015).
41. Plikus, M. V. et al. Regeneration of fat cells from myofibroblasts during wound healing. *Science* **355**, 748–752 (2017).
The ability of myofibroblasts to convert to adipocytes is unexpected, because the two cell types were thought to represent distinct lineages.
42. Mastrogiannaki, M. et al. β-Catenin stabilization in skin fibroblasts causes fibrotic lesions by preventing adipocyte differentiation of the reticular dermis. *J. Invest. Dermatol.* **136**, 1130–1142 (2016).
43. Michael, S., Achilleos, C., Panayiotou, T. & Strati, K. Inflammation shapes stem cells and stemness during infection and beyond. *Front. Cell Dev. Biol.* **4**, 118 (2016).
44. Karin, M. & Clevers, H. Reparative inflammation takes charge of tissue regeneration. *Nature* **529**, 307–315 (2016).
45. Aurora, A. B. & Olson, E. N. Immune modulation of stem cells and regeneration. *Cell Stem Cell* **15**, 14–25 (2014).
46. Shook, B., Xiao, E., Kumamoto, Y., Iwasaki, A. & Horsley, V. CD301b^{+} macrophages are essential for effective skin wound healing. *J. Invest. Dermatol.* **136**, 1885–1891 (2016).
47. Ali, N. et al. Regulatory T cells in skin facilitate epithelial stem cell differentiation. *Cell* **169**, 1119–1129 (2017).
48. Patel, A. S. et al. TIE2-expressing monocytes/macrophages regulate revascularization of the ischemic limb. *EMBO Mol. Med.* **5**, 858–869 (2013).
49. Cosin-Roger, J. et al. Hypoxia ameliorates intestinal inflammation through NLRP3/mTOR downregulation and autophagy activation. *Nat. Commun.* **8**, 98 (2017).
50. Jia, J. et al. LGR5 expression is controled by IKKα in basal cell carcinoma through activating STAT3 signaling pathway. *Oncotarget* **7**, 27280–27294 (2016).
51. Lindemans, C. A. et al. Interleukin-22 promotes intestinal-stem-cell-mediated epithelial regeneration. *Nature* **528**, 560–564 (2015).
52. Naik, S. et al. Inflammatory memory sensitizes skin epithelial stem cells to tissue damage. *Nature* **550**, 475–480 (2017).
Inflammation can trigger the epigenetic memory of injury by maintaining chromosomal accessibility to key stress-response genes.
53. Meng, S., Chanda, P., Thandavarayan, R. A. & Cooke, J. P. Transflammation: innate immune signaling in nuclear reprogramming. *Adv. Drug Deliv. Rev.* **120**, 133–141 (2017).
54. Ori, D., Murase, M. & Kawai, T. Cytosolic nucleic acid sensors and innate immune regulation. *Int. Rev. Immunol.* **36**, 74–88 (2017).
55. Yang, H., Adam, R. C., Ge, Y., Hua, Z. L. & Fuchs, E. Epithelial–mesenchymal micro-niches govern stem cell lineage choices. *Cell* **169**, 483–496 (2017).
56. Viswanathan, P. et al. Mimicking the topography of the epidermal-dermal interface with elastomer substrates. *Integr. Biol. (Camb.)* **8**, 21–29 (2016).
57. Miroshnikova, Y. A. et al. Adhesion forces and cortical tension couple cell proliferation and differentiation to drive epidermal stratification. *Nat. Cell Biol.* **20**, 69–80 (2018).
58. Gudipaty, S. A. et al. Mechanical stretch triggers rapid epithelial cell division through Piezo1. *Nature* **543**, 118–121 (2017).
59. Panciera, T. et al. Induction of expandable tissue-specific stem/progenitor cells through transient expression of YAP/TAZ. *Cell Stem Cell* **19**, 725–737 (2016).
60. Walko, G. et al. A genome-wide screen identifies YAP/WBP2 interplay conferring growth advantage on human epidermal stem cells. *Nat. Commun.* **8**, 14744 (2017).
61. Totaro, A. et al. YAP/TAZ link cell mechanics to Notch signalling to control epidermal stem cell fate. *Nat. Commun.* **8**, 15206 (2017).
62. Yui, S. et al. YAP/TAZ-dependent reprogramming of colonic epithelium links ECM remodeling to tissue regeneration. *Cell Stem Cell* **22**, 35–49 (2018).
63. Liu, X. et al. ROCK inhibitor and feeder cells induce the conditional reprogramming of epithelial cells. *Am. J. Pathol.* **180**, 599–607 (2012).
64. Wang, X. et al. Cloning and variation of ground state intestinal stem cells. *Nature* **522**, 173–178 (2015).
65. Mou, H. et al. Dual SMAD signaling inhibition enables long-term expansion of diverse epithelial basal cells. *Cell Stem Cell* **19**, 217–231 (2016).
66. Owens, D. M., Romero, M. R., Gardner, C. & Watt, F. M. Suprabasal α6β4 integrin expression in epidermis results in enhanced tumourigenesis and disruption of TGFβ signalling. *J. Cell Sci.* **116**, 3783–3791 (2003).
67. Schultz, M. B. & Sinclair, D. A. When stem cells grow old: phenotypes and mechanisms of stem cell aging. *Development* **143**, 3–14 (2016).
68. Avgustinova, A. & Benitah, S. A. Epigenetic control of adult stem cell function. *Nat. Rev. Mol. Cell Biol.* **17**, 643–658 (2016).
69. Tilly, J. L. & Sinclair, D. A. Germline energetics, aging, and female infertility. *Cell Metab.* **17**, 838–850 (2013).
70. Tomasetti, C. & Vogelstein, B. Variation in cancer risk among tissues can be explained by the number of stem cell divisions. *Science* **347**, 78–81 (2015).
71. Yang, T. B. et al. Mutual reinforcement between telomere capping and canonical Wnt signalling in the intestinal stem cell niche. *Nat. Commun.* **8**, 14766 (2017).
72. Conboy, I. M. et al. Rejuvenation of aged progenitor cells by exposure to a young systemic environment. *Nature* **433**, 760–764 (2005).
73. Rebo, J. et al. A single heterochronic blood exchange reveals rapid inhibition of multiple tissues by old blood. *Nat. Commun.* **7**, 13363 (2016).
74. Neves, J., Sousa-Victor, P. & Jasper, H. Rejuvenating strategies for stem cell-based therapies in aging. *Cell Stem Cell* **20**, 161–175 (2017).
75. Zhang, Y. et al. Development and stem cells of the esophagus. *Semin. Cell Dev. Biol.* **66**, 25–35 (2017).
76. Matsumura, H. et al. Hair follicle aging is driven by transepidermal elimination of stem cells via COL17A1 proteolysis. *Science* **351**, aad4395 (2016).
77. Watanabe, M. et al. Type XVII collagen coordinates proliferation in the interfollicular epidermis. *eLife* **6**, e26635 (2017).
78. Nalapareddy, K. et al. Canonical Wnt signaling ameliorates aging of intestinal stem cells. *Cell Rep.* **18**, 2608–2621 (2017).
79. Scudellari, M. To stay young, kill zombie cells. *Nature* **550**, 448–450 (2017).
80. Latil, M. et al. Cell-type-specific chromatin states differentially prime squamous cell carcinoma tumor-initiating cells for epithelial to mesenchymal transition. *Cell Stem Cell* **20**, 191–204 (2017).
81. Brown, S. et al. Correction of aberrant growth preserves tissue homeostasis. *Nature* **548**, 334–337 (2017).
Demonstration that wild-type cells are involved in actively eliminating cells with an activating β-catenin mutation from the epidermis.
82. Radyk, M. D., Burclaff, J., Willet, S. G. & Mills, J. C. Metaplastic cells in the stomach arise, independently of stem cells, via dedifferentiation or transdifferentiation of chief cells. *Gastroenterology* **154**, 839–843 (2018).
83. Jiang, M. et al. Transitional basal cells at the squamous–columnar junction generate Barrett's oesophagus. *Nature* **550**, 529–533 (2017).
84. Duarte, D. et al. Inhibition of endosteal vascular niche remodelling rescues hematopoietic stem cell loss in AML. *Cell Stem Cell* **22**, 64–77 (2017).
Pharmacologic manipulation of the haematopoietic stem cell niche to preserve normal function is a new paradigm for treating acute myeloid leukemia.
85. Mishra, A. et al. A protein phosphatase network controls the temporal and spatial dynamics of differentiation commitment in human epidermis. *eLife* **6**, e27356 (2017).
86. Haber, A. L. et al. A single-cell survey of the small intestinal epithelium. *Nature* **551**, 333–339 (2017).
87. Rubino, F. & Amiel, S. A. Is the gut the "sweet spot" for the treatment of diabetes? *Diabetes* **63**, 2225–2228 (2014).
88. Barker, N. et al. Lgr5^{+ve} stem cells drive self-renewal in the stomach and build long-lived gastric units in vitro. *Cell Stem Cell* **6**, 25–36 (2010).
89. Sigal, M. et al. Stromal R-spondin orchestrates gastric epithelial stem cells and gland homeostasis. *Nature* **548**, 451–455 (2017).
90. Arnold, K. et al. Sox2^{+} adult stem and progenitor cells are important for tissue regeneration and survival of mice. *Cell Stem Cell* **9**, 317–329 (2011).

91. Choi, E. et al. Lrig1$^+$ gastric isthmal progenitor cells restore normal gastric lineage cells during damage recovery in adult mouse stomach. Gut https://doi.org/10.1136/gutjnl-2017-313874 (2017).
92. Hayakawa, Y. et al. Mist1 expressing gastric stem cells maintain the normal and neoplastic gastric epithelium and are supported by a perivascular stem cell niche. *Cancer Cell* **28**, 800–814 (2015).
93. Wang, B., Zhao, L., Fish, M., Logan, C. Y. & Nusse, R. Self-renewing diploid Axin2$^+$ cells fuel homeostatic renewal of the liver. *Nature* **524**, 180–185 (2015). **This study showed that a subpopulation of Wnt-responsive hepatocytes broadly contributes to normal liver maintenance.**
94. Stanger, B. Z. Cellular homeostasis and repair in the mammalian liver. *Annu. Rev. Physiol.* **77**, 179–200 (2015).
95. Zorn, A. M. & Wells, J. M. Vertebrate endoderm development and organ formation. *Annu. Rev. Cell Dev. Biol.* **25**, 221–251 (2009).
96. Simmini, S. et al. Transformation of intestinal stem cells into gastric stem cells on loss of transcription factor Cdx2. *Nat. Commun.* **5**, 5728 (2014). **This study demonstrated that one gene, *Cdx2*, functionally distinguishes between intestinal stem cells and gastric stem cells.**
97. Wells, J. M. Developmental biology: regional identity of gut stem cells—one gene to rule them all. *Nat. Rev. Gastroenterol. Hepatol.* **12**, 125–126 (2015).
98 Takahashi, K. &Yamanaka, S. Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell* **126**, 663–676 (2006).

Acknowledgements F.M.W. acknowledges funding from the Wellcome Trust (206439/Z/17/Z), Medical Research Council (MR/PO18823/1), Cancer Research UK (C219/A23522) and the Biotechnology and Biological Sciences Research Council (BB/M007219/1). J.M.W. is supported by grants from the National Institutes of Health R01DK092456, U19AI116491, P01HD093363 and U01DK103117. We thank C. Mooney for help with the Figures.

Reviewer information Nature thanks C. Lengner and J. Rajagopal for their contribution to the peer review of this work.

Author contributions J.M.W. and F.M.W. conceived and wrote the manuscript and drafted the Figures together.

Competing interests The authors declare no competing interests.

Additional information

Reprints and permissions information is available at http://www.nature.com/reprints.

Correspondence and requests for materials should be addressed to J.M.W. or F.M.W.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

https://doi.org/10.1038/s41586-018-0087-1

New approaches for brain repair—from rescue to reprogramming

Roger A. Barker[1]*, Magdalena Götz[2,3]* & Malin Parmar[4]*

The ability to repair or promote regeneration within the adult human brain has been envisioned for decades. Until recently, such efforts mainly involved delivery of growth factors and cell transplants designed to rescue or replace a specific population of neurons, and the results have largely been disappointing. New approaches using stem-cell-derived cell products and direct cell reprogramming have opened up the possibility of reconstructing neural circuits and achieving better repair. In this Review we briefly summarize the history of neural repair and then discuss these new therapeutic approaches, especially with respect to chronic neurodegenerative disorders.

The mammalian central nervous system (CNS), unlike many other organs, has a limited capacity for self-repair, and the ability to overcome this limitation would transform how we could treat patients with a vast array of neurological disorders. The relative lack of neurogenesis, failure of axonal regeneration and non-permissive local environment in the adult CNS have been major—and literal—barriers to CNS repair. Nevertheless, in recent years it has become clear that much can be done in the mature CNS to foster innate regeneration.

Diseases or disorders of the CNS can be developmental, inherited or acquired and acute or chronic, and each presents a different set of challenges when developing repair strategies. Both acute damage and chronic neurodegenerative disorders are characterized by an inter-related macroglial and microglial/inflammatory reaction that involves changes in the extracellular matrix and/or alterations in the blood–brain barrier. Therefore, agents that improve this environment could be used to help to protect neurons from further damage and to promote circuit reformation and function. Such disease-modifying therapies—using, for example, trophic factors or cell transplants that release such factors—could in theory be used alone in the early stages of disease or combined with approaches designed to replace lost cells through cell transplantation, re-activation of endogenous neurogenesis or directed reprogramming in situ. In this Review, we discuss three main strategies for restoring function in the damaged or diseased CNS: (i) cell rescue using neurotrophic factors or cell-based approaches with putative-disease modifying effects; (ii) cell replacement therapies using transplants of exogenously derived and in vitro cultured cells; and (iii) cell replacement using endogenous neural precursor cells or strategies that involve the direct reprogramming of resident cells within the brain (Fig. 1).

Neurotrophic factors and disease modification

Enhancing the regenerative potential of the CNS in the face of a disease process can be done in two major ways. The first is to use approaches that restore and maintain the functional integrity of cells that are dysfunctional but not lost. This typically involves neurotrophic factors, some of which could be delivered using cell-based treatments. However, grafted cells could also be used to protect dysfunctional cells from ongoing insults such as metabolic stress. The second strategy is to minimize environmental barriers to repair, for example, by modifying the extracellular matrix and/or glial and immune reactions[1,2].

Exercise-based therapies and cognitive training have been shown experimentally to promote endogenous neurogenesis and the local production of growth factors such as BDNF in the brain[3]. Whether these changes in experimental animals can translate into clinical benefits is largely unknown, but such interventions do have the potential to be used in combination with cell replacement or cell rescue therapies to optimize their reparative benefits.

The use of trophic factors has been explored for over 20 years, most notably in amyotrophic lateral sclerosis (ALS) and Parkinson's disease[4,5] (Table 1). These initial studies suffered from major problems with delivery—the factors either failed to reach their neuronal targets or were rapidly inactivated. This led to a new round of trials using intraparenchymal delivery, most notably involving delivery of GDNF into the striata of patients with moderately advanced Parkinson's disease. In two open-label studies, from two independent centres, this produced positive clinical responses, with post mortem and imaging data supporting its mode of action[6–9]. However, a subsequent double-blind placebo-controlled trial in 34 patients with Parkinson's disease found that patients treated with GDNF showed no significant benefit over placebo-treated patients after 6 months[10].

The reasons for the discordant results from these trials have been discussed extensively[11], including possible placebo effects, although these would not explain the reported changes on PET imaging and at post mortem in patients treated with GDNF infusions. Furthermore, in the double-blind study, the group treated with saline did not improve, which also argues against any placebo effects and suggests that the trial failed because the active treatment did not work. This could have been due to the dose of GDNF and/or the way in which it was delivered, both of which differed from those used in the open-label studies. Given this uncertainty, further trials have been undertaken, including a double-blind approach with a new convection-enhanced delivery system that delivers GDNF more effectively. However, so far positive outcomes have not been achieved (https://scienceofparkinsons.com/2016/07/07/the-gdnf-trial-bristol-initial-results/), which suggests that the use of trophic factors alone may not work to treat such chronic neurodegenerative disorders. This is reinforced by parallel studies involving the delivery of other GDNF-like factors[12], such as neurturin using an adeno-associated virus (AAV) delivery system (AAV-NTN). Here, the initial open-label studies showed efficacy but two subsequent

---

[1] Department of Clinical Neuroscience and Cambridge Stem Cell Institute, University of Cambridge, Cambridge, UK. [2] Institute for Stem Cell Research, Helmholtz Center Munich and Physiological Genomics, Biomedical Center, Ludwig-Maximilians University, Planegg-Martinsried, Germany. [3] SyNergy, Excellence Cluster Systems Neurology, Center for Stroke and Dementia, Munich, Germany. [4] Wallenberg Neuroscience Center and Lund Stem Cell Center, Lund University, Lund, Sweden. *e-mail: rab46@cam.ac.uk; magdalena.goetz@helmholtz-muenchen.de; malin.parmar@med.lu.se

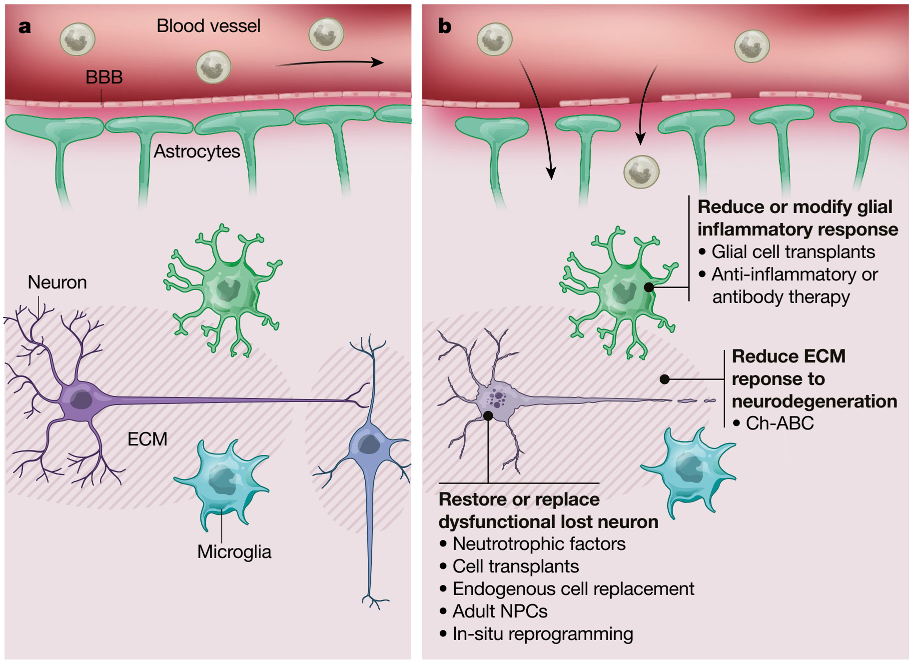


Fig. 1 | Strategies to repair the diseased brain. a, In the healthy brain, neurons project and form synapses with target neurons. Axons are supported by myelinating oligodendrocytes (not shown) and astrocytes, which form a syncytium that helps to protect neurons and maintain the blood–brain barrier (BBB). **b**, In disease states, such as those associated with neurodegenerative disorders, neurons become dysfunctional and die, and this is often accompanied by a glial reaction and a change in the ECM. Treatments to repair and restore the brain in such conditions could include: (i) rescuing cells using neurotrophic factors; (ii) cell transplants to replace lost cells or support those targeted by the disease process; (iii) recruitment of endogenous cells to replace those lost to the disease process; (iv) modification of the glial or inflammatory response; and/or (v) modification of the ECM.

double-blind placebo control studies (which differed in terms of where, and how, the AAV-NTN was delivered) failed to show efficacy[13–15].

So, does this mean that repairing the brain with trophic factors is not a viable way forward? This is likely to be the case when the disease has advanced beyond a certain stage; once a patient has lost a critical number of cells or fibres, it is unrealistic to expect any growth factor to work, as there is nothing left for it to regenerate. Given that most neurodegenerative disorders present clinically when there has already been extensive cell or fibre loss, the early use of these agents will be required, and a post hoc analysis of some of the data from GDNF and AAV-NTN trials supports this view[14,15]. Moreover, the disease state may also interfere with growth factor function. For example, GDNF may not be able to act through its normal receptor in the presence of α-synuclein pathology in Parkinson's disease because the pathology causes alterations in NURR1 signalling, which is needed for GDNF to produce its therapeutic effect[16]. Thus, it may be necessary to enhance levels of NURR1 before using GDNF to treat Parkinson's disease for it to work efficiently, necessitating the adoption of combination approaches in any future trials. Finally, the optimal dose of any growth factor and how to best deliver it, including the volume of distribution needed to maximise its benefits, are still important questions. To date, GDNF and AAV-NTN have been delivered in ways that are insufficient to cover the structure being treated, and many of the failures seen clinically may relate to this. Post mortem data from both the original GDNF and subsequent AAV-NTN studies provide some support for this idea[9,15].

Overcoming environmental hurdles to regeneration

Promoting neuronal survival and outgrowth is only one aspect of the problem in repairing the CNS. The other is enhancing the environment to support that regenerative process. This essentially distils down to modifying the glial elements and/or the extracellular matrix (ECM), both of which have been reported to produce signals that inhibit repair in the adult CNS[1,2].

In terms of the glial cells, it has long been known that both astrocytes and oligodendrocytes express cell surface molecules that inhibit regenerating axons, although the expression and extent of these varies as a function of disease state (for example, being greater in patients with acute traumatic injuries than in chronic neurodegenerative disorders). The premise that glial cells produce inhibitory factors that can be blocked has led to early clinical trials, most notably using antibodies against the myelin-derived molecule NOGO in patients with spinal cord injury (http://www.sci-research.uzh.ch/en/research/ClinicalTrials/Nogo.html). However, the concept that glia are 'bad' for CNS repair has been challenged, in that the glial scar or glial reaction to injury or disease involves a heterogeneous collection of cells, which have a variety of functions. So, under some conditions the glial response may be beneficial for repair, possibly through effects on the inflammatory response to CNS disease[1]. Conversely under other circumstances, the opposite may occur. For example, it has recently been reported that microglia can induce astrocytes to become neurotoxic[17]. This concept of 'gliopathy' is not new and has been explored in models of ALS, where transplanted glial cells have been used to buffer extracellular glutamate and thus protect motor neurons[18,19]. Whether such an approach could be used to treat ALS clinically is unknown, but this work highlights the potential use of non-neuronal cell transplants to protect neurons in disease states.

More recently there has been interest in the ECM of the brain, and especially peri-neuronal nets (PNNs). PNNs are unevenly distributed across the brain and spinal cord and appear to have many functions during development and in the adult brain, including a role in neural plasticity. This role has been shown to extend to neurodegenerative disorders, most notably in transgenic models of Alzheimer's disease, where removal of PNNs, either using chrondoitinase-ABC or genetically, has been shown to improve cognitive function without influencing the underlying disease process, possibly by enhancing neurite outgrowth and synaptic plastcity[20,21]. These results suggest that the use of this, or similar, agents may have benefits when combined with cell therapies in which synaptic integration with host cells may be restricted by local PNNs and/or a disease-driven increase in expression of inhibitory ECM molecules.

Replacement therapies using exogenous cells

When a neurodegenerative disorder or injury reaches the point at which a critical number of neurons has been lost, the replacement of these neurons is essential for functional improvement. Lost cells could be replaced by transplantation into single or multiple brain regions to partially recreate complex circuits either locally or over long distances.

For this to work, a number of important questions need to be addressed. First, what is the optimal cell type(s) to be transplanted and at what developmental stage should they be collected for grafting? Second, what CNS location(s) are needed for optimal recovery and fibre outgrowth? Third, how many cells are needed for effective repair, and finally, should other interventions be done either at the same time as grafting or after transplantation to enhance survival, integration or function of the cells?

Table 1 | Clinical trials using cell therapies or growth factors in patients with different neurodegenerative disorders

Neurodegenerative disease	Use of neurotrophic factors in clinical trials	Use of cell transplants in clinical trials
Alzheimer's disease	NGF	No
Parkinson's disease	NGF GDNF/NTN	Human fetal ventral midbrain tissue Adrenal medulla Carotid body Retinal pigmentary epithelial cells (spheramine)
Amyotrophic lateral sclerosis	CNTF BDNF IGF1	Ongoing (neural precursor cells delivering GDNF)
Huntington's disease	CNTF	Human fetal striatal tissue

Cell replacement therapy using human fetal allografts

Parkinson's disease is one of the least complex neurodegenerative disorders to target with cell-based therapies, because local degeneration of dopamine neurons in the midbrain is critical for the clinical expression of many features of this disorder. It has therefore been the subject of a relatively large number of clinical transplantation trials using developing midbrain dopamine cells derived from the human fetal ventral midbrain[22]. Although the outcome of these trials has been highly variable[23], they have shown that at least a sub-group of patients with Parkinson's disease that receive such transplants at different centres do well[24–27]. However, the results have also shown that, over time, a minority of the transplanted cells become affected by the disease process and show protein aggregation[28,29], and that some patients can experience troublesome side effects such as graft-induced dyskinesias[30–32] relating to graft composition and capacity to innervate the whole putamen[22].

Clinical trials using transplants of primary human fetal striatal tissue have also been performed in patients with Huntington's disease[33–35], on the premise that the medium spiny neurons of the striatum are the neurons that are primarily lost in this disease[36]. In this case, the neurons are placed homotopically into the striatum to reconstruct basal ganglia circuits. To date the effects in patients have been modest at best[33–35], with evidence for poor graft survival and integration[37].

Other neurodegenerative and CNS disorders have more complex pathologies, involving several cell types and/or cells in many regions of the CNS, and it is hard to see how these could be treated using fetal tissue transplants. However, stem cells may be a more promising source of cells for transplantation in such disorders, because they can be differentiated into a wide range of neuronal phenotypes.

Cell replacement therapy using stem cells

A major challenge for developing any cell replacement therapy is the ability to reliably and robustly generate large numbers of relevant cells. Recent developments using human embryonic stem cells (hES cells) and induced pluripotent stem cells (iPS cells) have now made this possible[38] and clinical trials using this approach are imminent, for example, in patients with Parkinson's disease[39,40].

Stem-cell-derived dopamine neurons have also been used to better understand graft function, innervation and integration in pre-clinical animal models of Parkinson's disease. For example, comparative studies with human fetal ventral midbrain tissue have shown that both hESC and hiPSC-derived dopaminergic progenitors are very similar to their fetal counterparts[41,42] and mediate functional recovery with equivalent potency over a similar time course[43]. New methodologies have also been used to study their mechanism of action, as well as their ability to appropriately innervate and integrate into host circuitry. These studies have shown that both fetal and stem-cell-derived dopaminergic neurons transplanted to the rat brain can innervate the correct target structures when placed either locally at the site of action (in the striatum) or back in the midbrain, which requires more extensive axonal outgrowth from the cells[42,43]. Furthermore, these cells not only innervate but also integrate into host neuronal circuitry[44], and produce functional recovery[38,45]. Similarly, transplantation of stem-cell-derived medium spiny neurons have now been shown to survive and function in rodent models of striatal degeneration (used to model Huntington's disease)[46–48].

Another advantage of using stem-cell-derived transplants is the ability to remove or minimize unwanted neuronal subtypes; for example, serotonergic neurons that may mediate the graft-induced dyskinesias seen in patients that have received transplants of human fetal ventral midbrain tissue to treat Parkinson's disease. Furthermore, it is possible to study the effect of grafting pure populations of neurons, or a combination of neuronal and or astrocytic cell types. Indeed, deficient astrocyte function may contribute to the loss of neurons in many neurodegenerative disorders, and the absence of such cells may affect neuronal survival in transplants, as has been shown in fetal striatal grafts in patients with Huntington's disease[49]. As an alternative to possible co-grafting with glial cells, other agents or biomaterials engineered to express growth factors, immune-modulating components or similar substances to provide a supportive and protective microenvironment could be delivered together with the transplanted neurons; this could even include some form of scaffold for larger lesions such as stroke[50].

Genetic modifications or correction of cells could also be used before transplantation to make the grafted cells resistant or refractory to the disease pathology. This may be of critical importance when the transplants are derived from the patients' own cells[42], or in diseases such as ALS, where motor neuron loss is driven in part by astrocytes. However, even unrelated healthy donor cells can be affected by proteinopathies in some neurodegenerative disorders, as has been seen in trials of fetal tissue transplantation for Parkinson's disease and Huntington's disease[28,29,51]. In these cases, making cells resistant to the disease by removing the normal protein onto which the pathological species templates could be helpful.

Additionally, as genetic engineering tools become more refined, stem cells could be modified to enhance aspects of their behaviour. Such techniques could include modifying chemotactic interactions between the grafted cells and their progeny to enhance migration[52]; making cells refractory to the inhibitory cues of the glial scar; increasing their innervation potential by enhanced expression of polysialic acid[53] (all of which would need to done so that the enhanced outgrowth was targeted specifically to the area for repair); and/or decreasing their immunogenicity[54,55]. Based on cell transplants in animal models, the possibility of engineering stem cell progeny so that their function can be modified using optogenetic or chemogenetic interventions is now also being considered[45,56,57].

Cell replacement therapy for circuit reconstruction

Owing to the complexity of the organization of the cerebral cortex, replacing neurons in this part of the brain is a more challenging task, and is thus still at the pre-clinical stage. New technologies for studying cell integration and function have shown that transplanted cells mature, integrate into local host circuitry, form functional synapses, mediate recovery and provide circuit repair in animal models of Alzheimer's disease, stroke and cortical lesions[58–62]. Notably, it has recently been shown that the exact brain-wide input connectome of neurons in the mouse visual cortex can be replicated by transplanted neurons[59]. This is important, as aberrant connectivity could lead to deleterious consequences such as epileptic seizures. Thus, the adult mammalian brain is more plastic than previously thought in terms of accommodating new neurons and building new functional circuitry.

Replacement therapies using endogenous cells

Brain repair from endogenous sources has been considered for some time[63], especially after the (re-)discovery of adult neurogenesis and proof of its existence in the human brain[64,65], prompting attempts to recruit new neurons from endogenous neurogenic niches to achieve repair in the injured brain[66,67] (Fig. 2). Simply recruiting new neurons from active neurogenic sites appears intuitively to be the most promising approach, and the easiest to achieve. However, in most cases, the subtypes of neurons needed for repair are different from those generated endogenously. Hence, attempts to recruit neuroblasts to form neurons at other sites have met with only limited success, with many of the neuroblasts streaming to the injury site but then failing to fully differentiate into appropriate neuronal types that can survive in the long term[67,68] (Fig. 2). Adult neurogenesis is very species-specific[69], though, so endogenous striatal neurogenesis that has been said to occur in the human (but not rodent)[70,71] brain may prove useful and be more easily manipulated to generate the relevant neuronal subtypes lost in the striatum after stroke or in neurodegenerative disorders[72,73].

A further approach to minimize neuronal loss or help repair and replacement is through the recruitment of stem-cell-derived astrocytes. These differ from the local parenchymal astrocytes and release an array of beneficial growth factors[74]. However, the extent to which this occurs in human patients and can be used for therapy is currently unknown.

Excitingly, attempts to turn stem-cell-derived astrocytes or other parenchymal glial cells into functional neurons by forced expression of neurogenic fate determinants have progressed rapidly. This idea was first tested in the nervous system in 2002, by turning glial cells into neurons in vitro[75]. The first attempts to achieve this in an injured brain in vivo in 2005 were exciting as proof-of-principle experiments, but disappointing in terms of quantitative outcome[76–78]. However, outcomes have improved recently[79–81], and there is now encouraging data on fate conversion of local non-neuronal cells towards a neuronal phenotype of the sort affected by the disease process.

A single neurogenic transcription factor (acting also as potent fate determinant in development) can in many cases be sufficient to convert glial cells into fully functional neurons[82]. For example, neurogenin 2 or NEUROD1 can produce glutamatergic neurons from proliferating glial cells in the adult cerebral cortex, including in the highly inflamed brain environment found after traumatic brain injury[79,83,84]. The key to generating highly efficient neuronal conversion protocols is not necessarily to add more transcription factors, but rather to keep the induced neurons alive by protecting them from death and promoting their maturation. Many neurons induced by direct reprogramming from glial cells die because of the local generation of excessive reactive oxygen species (ROS), in a process known as ferroptosis[79,84]. By keeping the new neurons alive and reducing ROS levels, conversion efficiencies of over 90% have been achieved, with almost all glial cells, transduced with the neurogenic transcription factor, acquiring a neuronal identity within 1–3 weeks[84]. Importantly, these neurons also acquired a pyramidal neuronal morphology with the identity of deep layer neurons, at least at the transcriptional level. This approach has also shown promise in other brain areas, including in the minimally injured striatum (using SOX2 and growth factor treatment), the midbrain and the spinal cord[85], and in the cerebral cortex using NEUROD1 and an amyloid plaque deposition model[83]. In these latter approaches, a mix of glutamatergic or GABAergic (γ-aminobutyric-acid-releasing) neurons has been obtained, which may be beneficial if several types of neurons need to be replaced or synaptic plasticity increased by providing new GABAergic interneurons, as shown through transplantation studies[86,87]. Synaptic plasticity and new circuit formation using these approaches could also be further promoted by co-administration of rehabilitation therapies (see above).

The promising outcome of direct reprogramming approaches raises two important issues: which glial cells to target and which vectors to use. Targeting proliferating glial cells has obvious advantages[88], but they are heterogeneous[1,76,89]. Specific targeting of glial subtypes has been achieved using subtype-specific promoters that either drive reprogramming factors directly[83,90,91] or drive cell-type-specific Cre expression in transgenic mice[80,81,85]. This approach has now been shown to work with AAVs that not only have been used in patients clinically but also can preferentially infect glial cells[92–94]. AAVs have been used to reprogram oligodendrocyte progenitor cells[80,95] to a fate that is dependent on the transcription factor combination—for example, generating fast-spiking parvalbumin-expressing interneurons in the intact striatum[95]. Moreover, these neurons have been shown to receive input from local interneurons[80]. Finally, recent evidence suggests that striatal and midbrain astrocytes and NG2 glia can be converted to neurons with functional GABAergic- and dopaminergic-like phenotypes in a dopamine depletion model[96]. While these pioneering studies show that newly induced neurons can connect locally and result in behavioural changes, their long-distance inputs and outputs and contribution to the neuronal network still remain to be explored.

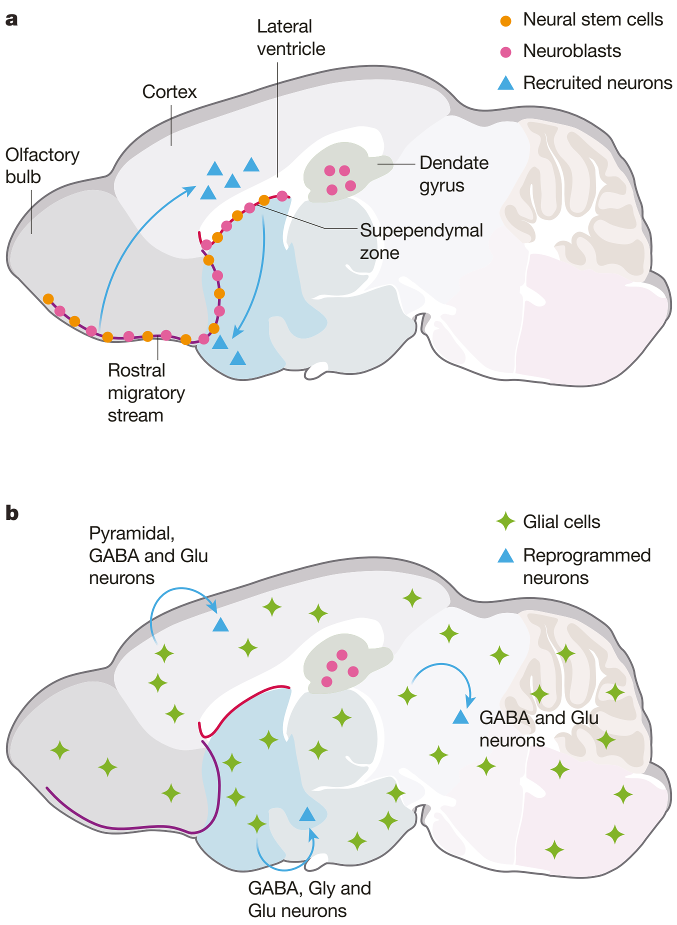


Fig. 2 | Strategies for in situ repair. a, Recruitment of new neurons from endogeneous precursors originating from sites of adult neurogenesis. **b**, Directed reprogramming of resident non-neuronal, glial cells.

Direct neuronal reprogramming in the adult brain has come a long way in the past few years, but it remains to be seen how these exciting new approaches will translate to the clinic. While there are no major obstacles to turning adult human cells into functioning subtype-specific neurons, as has been shown, for example, with fibroblasts or pericytes in vitro[97–100], neuronal reprogramming in vivo in the diseased brain has proven more challenging[79]. Thus, investigation of differences in the starting cells that affect the outcome of neuronal reprogramming, as well as influences in the local (disease) environment, is one of the next requirements to further this approach[79]. Moreover, tools need to be developed for non-invasive reprogramming techniques, either using systemic application of viral vectors (as is the case for some AAVs[92–94]) or using small molecules that have been shown to efficiently achieve direct neuronal reprogramming in vitro[79,101].

Direct reprogramming also brings new standards to the field, as it is important to compare the induced cell type to the endogenous counterpart that is being replaced or instructed (for example, iPS cells to ES cells). This can be done using functional assessments and genome-wide expression analysis. Such techniques are urgently needed for the field of neuronal replacement, as it is so far unknown how authentic the transplanted or reprogrammed neurons really are, compared to their endogenous counterparts. Raising these standards in basic research and then translating them to the clinic will be a challenge for the future.

Conclusion and future perspectives

A broad approach will be needed to deliver effective repair strategies to the human brain that combine established approaches with new technologies, and ultimately may involve combination therapies. This progress will be possible only if we better understand disease states and the contribution of all cells and the ECM to the processes of damage and cell loss, as well as the innate regenerative ability of the adult brain and its capacity to be reprogrammed. By combining approaches to optimize the disease environment and synaptic plasticity while providing new neurons for replacement therapy, new circuits could be rebuilt with functional benefits across a range of disorders, and this could eventually develop into a more standard treatment. However, this will need to be an iterative approach with an ongoing dialogue between preclinical work and clinical trials. It will also be crucial to avoid the temptation to short-circuit this approach with premature claims and commercialization, which have the potential to derail this whole field.

Received: 18 January 2018; Accepted: 15 March 2018;
Published online 16 May 2018.

1. Adams, K. L. & Gallo, V. The diversity and disparity of the glial scar. *Nat. Neurosci.* **21**, 9–15 (2018).
2. Fawcett, J. W. The extracellular matrix in plasticity and regeneration after CNS injury and neurodegenerative disease. *Prog. Brain Res.* **218**, 213–226 (2015).
3. Chieffi, S. et al. Neuroprotective effects of physical activity: evidence from human and animal studies. *Front. Neurol.* **8**, 188 (2017).
4. Bartus, R. T. & Johnson, E. M. Jr. Clinical tests of neurotrophic factors for human neurodegenerative diseases, part 1: where have we been and what have we learned? *Neurobiol. Dis.* **97**, 156–168 (2017).
5. Bartus, R. T. & Johnson, E. M. Jr. Clinical tests of neurotrophic factors for human neurodegenerative diseases, part 2: where do we stand and where must we go next? *Neurobiol. Dis.* **97**, 169–178 (2017).
6. Gill, S. S. et al. Direct brain infusion of glial cell line-derived neurotrophic factor in Parkinson disease. *Nat. Med.* **9**, 589–595 (2003).
7. Patel, N. K. et al. Intraputamenal infusion of glial cell line-derived neurotrophic factor in PD: a two-year outcome study. *Ann. Neurol.* **57**, 298–302 (2005).
8. Slevin, J. T. et al. Unilateral intraputamenal glial cell line-derived neurotrophic factor in patients with Parkinson disease: response to 1 year of treatment and 1 year of withdrawal. *J. Neurosurg.* **106**, 614–620 (2007).
9. Love, S. et al. Glial cell line-derived neurotrophic factor induces neuronal sprouting in human brain. *Nat. Med.* **11**, 703–704 (2005).
10. Lang, A. E. et al. Randomized controlled trial of intraputamenal glial cell line-derived neurotrophic factor infusion in Parkinson disease. *Ann. Neurol.* **59**, 459–466 (2006).
11. Barker, R. A. Continuing trials of GDNF in Parkinson's disease. *Lancet Neurol.* **5**, 285–286 (2006).
12. Runeberg-Roos, P. et al. Developing therapeutically more efficient Neurturin variants for treatment of Parkinson's disease. *Neurobiol. Dis.* **96**, 335–345 (2016).
13. Marks, W. J. Jr et al. Gene delivery of AAV2-neurturin for Parkinson's disease: a double-blind, randomised, controlled trial. *Lancet Neurol.* **9**, 1164–1172 (2010).
14. Warren Olanow, C. et al. Gene delivery of neurturin to putamen and substantia nigra in Parkinson disease: a double-blind, randomized, controlled trial. *Ann. Neurol.* **78**, 248–257 (2015).
15. Bartus, R. T. et al. Post-mortem assessment of the short and long-term effects of the trophic factor neurturin in patients with α-synucleinopathies. *Neurobiol. Dis.* **78**, 162–171 (2015).
16. Decressac, M. et al. α-Synuclein-induced down-regulation of Nurr1 disrupts GDNF signaling in nigral dopamine neurons. *Sci. Transl. Med.* **4**, 163ra156 (2012).
17. Liddelow, S. A. et al. Neurotoxic reactive astrocytes are induced by activated microglia. *Nature* **541**, 481–487 (2017).
18. Das, M. M. et al. Human neural progenitors differentiate into astrocytes and protect motor neurons in aging rats. *Exp. Neurol.* **280**, 41–49 (2016).
19. Lepore, A. C. et al. Focal transplantation-based astrocyte replacement is neuroprotective in a model of motor neuron disease. *Nat. Neurosci.* **11**, 1294–1301 (2008).
20. Yang, S. et al. Antibody recognizing 4-sulfated chondroitin sulfate proteoglycans restores memory in tauopathy-induced neurodegeneration. *Neurobiol. Aging* **59**, 197–209 (2017).
21. Bradbury, E. J. et al. Chondroitinase ABC promotes functional recovery after spinal cord injury. *Nature* **416**, 636–640 (2002).
22. Barker, R. A., Drouin-Ouellet, J. & Parmar, M. Cell-based therapies for Parkinson disease—past insights and future potential. *Nat. Rev. Neurol.* **11**, 492–503 (2015).
23. Barker, R. A., Barrett, J., Mason, S. L. & Björklund, A. Fetal dopaminergic transplantation trials and the future of neural grafting in Parkinson's disease. *Lancet Neurol.* **12**, 84–91 (2013).
A useful summary of all the data from trials of fetal ventral midbrain transplantation in patients with Parkinson's disease over the last 25 years and their conclusions.
24. Li, W. et al. Extensive graft-derived dopaminergic innervation is maintained 24 years after transplantation in the degenerating parkinsonian brain. *Proc. Natl Acad. Sci. USA* **113**, 6544–6549 (2016).
A key study showing good graft survival and transplantation in a patient with Parkinson's disease 24 years after transplantation.
25. Kefalopoulou, Z. et al. Long-term clinical outcome of fetal cell transplantation for Parkinson disease: two case reports. *JAMA Neurol.* **71**, 83–87 (2014).
26. Hauser, R. A. et al. Long-term evaluation of bilateral fetal nigral transplantation in Parkinson disease. *Arch. Neurol.* **56**, 179–187 (1999).
27. Mendez, I. et al. Cell type analysis of functional fetal dopamine cell suspension transplants in the striatum and substantia nigra of patients with Parkinson's disease. *Brain* **128**, 1498–1510 (2005).
28. Li, J. Y. et al. Lewy bodies in grafted neurons in subjects with Parkinson's disease suggest host-to-graft disease propagation. *Nat. Med.* **14**, 501–503 (2008).
29. Kurowska, Z. et al. Signs of degeneration in 12–22-year old grafts of mesencephalic dopamine neurons in patients with Parkinson's disease. *J. Parkinsons Dis.* **1**, 83–92 (2011).
30. Politis, M. et al. Serotonergic neurons mediate dyskinesia side effects in Parkinson's patients with neural transplants. *Sci. Transl. Med.* **2**, 38ra46 (2010).
31. Freed, C. R. et al. Transplantation of embryonic dopamine neurons for severe Parkinson's disease. *N. Engl. J. Med.* **344**, 710–719 (2001).
32. Olanow, C. W. et al. A double-blind controlled trial of bilateral fetal nigral transplantation in Parkinson's disease. *Ann. Neurol.* **54**, 403–414 (2003).
33. Bachoud-Lévi, A. et al. Safety and tolerability assessment of intrastriatal neural allografts in five patients with Huntington's disease. *Exp. Neurol.* **161**, 194–202 (2000).
34. Barker, R. A. et al. The long-term safety and efficacy of bilateral transplantation of human fetal striatal tissue in patients with mild to moderate Huntington's disease. *J. Neurol. Neurosurg. Psychiatry* **84**, 657–665 (2013).
35. Hauser, R. A. et al. Bilateral human fetal striatal transplantation in Huntington's disease. *Neurology* **58**, 687–695 (2002).
36. Peschanski, M., Cesaro, P. & Hantraye, P. Rationale for intrastriatal grafting of striatal neuroblasts in patients with Huntington's disease. *Neuroscience* **68**, 273–285 (1995).
37. Cisbani, G. & Cicchetti, F. The fate of cell grafts for the treatment of Huntington's disease: the post-mortem evidence. *Neuropathol. Appl. Neurobiol.* **40**, 71–90 (2014).
38. Steinbeck, J. A. & Studer, L. Moving stem cells to the clinic: potential and limitations for brain repair. *Neuron* **86**, 187–206 (2015).
39. Barker, R. A., Parmar, M., Studer, L. & Takahashi, J. Human trials of stem cell-derived dopamine neurons for Parkinson's disease: dawn of a new era. *Cell Stem Cell* **21**, 569–573 (2017).
40. Abbott, A. Fetal-cell revival for Parkinson's. *Nature* **510**, 195–196 (2014).
41. Kirkeby, A. et al. Generation of regionally specified neural progenitors and functional neurons from human embryonic stem cells under defined conditions. *Cell Reports* **1**, 703–714 (2012).
42. Kikuchi, T. et al. Human iPS cell-derived dopaminergic neurons function in a primate Parkinson's disease model. *Nature* **548**, 592–596 (2017).
An important study showing the long-term survival and functional efficacy of iPS-cell-derived dopamine cells (including from patients with Parkinson's disease) in non-human primate models of Parkinson's disease.
43. Grealish, S. et al. Human ESC-derived dopamine neurons show similar preclinical efficacy and potency to fetal neurons when grafted in a rat model of Parkinson's disease. *Cell Stem Cell* **15**, 653–665 (2014).
An key study showing that grafted stem-cell-derived dopamine neurons function equivalently to fetal cells in terms of maturation, innervation capacity, potency and function.
44. Grealish, S. et al. Monosynaptic tracing using modified rabies virus reveals early and extensive circuit integration of human embryonic stem cell-derived neurons. *Stem Cell Reports* **4**, 975–983 (2015).
45. Chen, Y. et al. Chemical control of grafted human PSC-derived neurons in a mouse model of Parkinson's disease. *Cell Stem Cell* **18**, 817–826 (2016).
46. Delli Carri, A. et al. Human pluripotent stem cell differentiation into authentic striatal projection neurons. *Stem Cell Rev.* **9**, 461–474 (2013).
47. Reidling, J. C. et al. Human neural stem cell transplantation rescues functional deficits in R6/2 and Q140 Huntington's disease mice. *Stem Cell Reports* **10**, 58–72 (2018).
48. Ma, L. et al. Human embryonic stem cell-derived GABA neurons correct locomotion deficits in quinolinic acid-lesioned mice. *Cell Stem Cell* **10**, 455–464 (2012).
49. Cisbani, G. et al. Striatal allografts in patients with Huntington's disease: impact of diminished astrocytes and vascularization on graft viability. *Brain* **136**, 433–443 (2013).
50. Kokaia, Z. & Lindvall, O. Stem cell repair of striatal ischemia. *Prog. Brain Res.* **201**, 35–53 (2012).
51. Cicchetti, F. et al. Mutant huntingtin is present in neuronal grafts in Huntington disease patients. *Ann. Neurol.* **76**, 31–42 (2014).
52. Ladewig, J., Koch, P. & Brüstle, O. Auto-attraction of neural precursors and their neuronal progeny impairs neuronal migration. *Nat. Neurosci.* **17**, 24–26 (2014).

53. Battista, D., Ganat, Y., El Maarouf, A., Studer, L. & Rutishauser, U. Enhancement of polysialic acid expression improves function of embryonic stem-derived dopamine neuron grafts in Parkinsonian mice. *Stem Cells Transl. Med.* **3**, 108–113 (2014).
54. Morizane, A. et al. MHC matching improves engraftment of iPSC-derived neurons in non-human primates. *Nat. Commun.* **8**, 385 (2017).
55. Gornalusse, G. G. et al. HLA-E-expressing pluripotent stem cells escape allogeneic responses and lysis by NK cells. *Nat. Biotechnol.* **35**, 765–772 (2017).
56. Aldrin-Kirk, P. et al. DREADD modulation of transplanted DA neurons reveals a novel Parkinsonian dyskinesia mechanism mediated by the serotonin 5-HT6 receptor. *Neuron* **90**, 955–968 (2016).
57. Steinbeck, J. A. et al. Optogenetics enables functional analysis of human embryonic stem cell-derived grafts in a Parkinson's disease model. *Nat. Biotechnol.* **33**, 204–209 (2015).
58. Espuny-Camacho, I. et al. Hallmarks of Alzheimer's disease in stem-cell-derived human neurons transplanted into mouse brain. *Neuron* **93**, 1066–1081 (2017).
59. Falkner, S. et al. Transplanted embryonic neurons integrate into adult neocortical circuits. *Nature* **539**, 248–253 (2016).
60. Espuny-Camacho, I. et al. Pyramidal neurons derived from human pluripotent stem cells integrate efficiently into mouse brain circuits in vivo. *Neuron* **77**, 440–456 (2013).
61. Tornero, D. et al. Synaptic inputs from stroke-injured brain to grafted human stem cell-derived neurons activated by sensory stimuli. *Brain* **140**, 692–706 (2017).
62. Michelsen, K. A. et al. Area-specific reestablishment of damaged circuits in the adult cerebral cortex by cortical neurons derived from mouse embryonic stem cells. *Neuron* **85**, 982–997 (2015).
63. Grade, S. & Götz, M. Neuronal replacement therapy: previous achievements and challenges ahead. *NPJ Regen. Med.* **2**, 29 (2017).
64. Frisén, J. Neurogenesis and gliogenesis in nervous system plasticity and repair. *Annu. Rev. Cell Dev. Biol.* **32**, 127–141 (2016).
65. Paredes, M. F., Sorrells, S. F., Garcia-Verdugo, J. M. & Alvarez-Buylla, A. Brain size and limits to adult neurogenesis. *J. Comp. Neurol.* **524**, 646–664 (2016).
66. Nakatomi, H. et al. Regeneration of hippocampal pyramidal neurons after ischemic brain injury by recruitment of endogenous neural progenitors. *Cell* **110**, 429–441 (2002).
67. Arvidsson, A., Collin, T., Kirik, D., Kokaia, Z. & Lindvall, O. Neuronal replacement from endogenous precursors in the adult brain after stroke. *Nat. Med.* **8**, 963–970 (2002). **An early study that provided important evidence for endogenous neurogenesis after brain injury.**
68. Thored, P. et al. Long-term neuroblast migration along blood vessels in an area with transient angiogenesis and increased vascularization after stroke. *Stroke* **38**, 3032–3039 (2007).
69. Lipp, H. P. & Bonfanti, L. Adult neurogenesis in mammals: variations and confusions. *Brain Behav. Evol.* **87**, 205–221 (2016).
70. Wang, C. et al. Human and monkey striatal interneurons are derived from the medial ganglionic eminence but not from the adult subventricular zone. *J. Neurosci.* **34**, 10906–10923 (2014).
71. Ernst, A. et al. Neurogenesis in the striatum of the adult human brain. *Cell* **156**, 1072–1083 (2014).
72. Grande, A. et al. Environmental impact on direct neuronal reprogramming in vivo in the adult brain. *Nat. Commun.* **4**, 2373 (2013).
73. Magnusson, J. P. et al. A latent neurogenic program in astrocytes regulated by Notch signaling in the mouse. *Science* **346**, 237–241 (2014).
This work shows the in vivo conversion of astrocytes into neurons solely by removing Notch signalling and that this conversion occurs in an interesting region-specific manner only in the striatum, showing that these astrocytes are more prone to neurogenesis than others.
74. Benner, E. J. et al. Protective astrogenesis from the SVZ niche after injury is controlled by Notch modulator Thbs4. *Nature* **497**, 369–373 (2013).
75. Heins, N. et al. Glial cells generate neurons: the role of the transcription factor Pax6. *Nat. Neurosci.* **5**, 308–315 (2002).
This work pioneered the approach of converting glial cells into neurons and showed the successful conversion of postnatal glial cells into neurons.
76. Buffo, A. et al. Expression pattern of the transcription factor Olig2 in response to brain injuries: implications for neuronal repair. *Proc. Natl Acad. Sci. USA* **102**, 18183–18188 (2005).
77. Ohori, Y. et al. Growth factor treatment and genetic manipulation stimulate neurogenesis and oligodendrogenesis by endogenous neural progenitors in the injured adult spinal cord. *J. Neurosci.* **26**, 11948–11960 (2006).
78. Kronenberg, G. et al. Modulation of fate determinants Olig2 and Pax6 in resident glia evokes spiking neuroblasts in a model of mild brain ischemia. *Stroke* **41**, 2944–2949 (2010).
79. Gascón, S., Masserdotti, G., Russo, G. L. & Götz, M. Direct neuronal reprogramming: achievements, hurdles, and new roads to success. *Cell Stem Cell* **21**, 18–34 (2017).
80. Torper, O. et al. In vivo reprogramming of striatal NG2 glia into functional neurons that integrate into local host circuitry. *Cell Rep.* **12**, 474–481 (2015).
This work shows the in vivo connectivity of reprogrammed neurons in the striatum that receive local, but not long-distance, inputs.
81. Torper, O. et al. Generation of induced neurons via direct conversion in vivo. *Proc. Natl Acad. Sci. USA* **110**, 7038–7043 (2013).
82. Masserdotti, G., Gascón, S. & Götz, M. Direct neuronal reprogramming: learning from and for development. *Development* **143**, 2494–2510 (2016).
83. Guo, Z. et al. In vivo direct reprogramming of reactive glial cells into functional neurons after brain injury and in an Alzheimer's disease model. *Cell Stem Cell* **14**, 188–202 (2014).
84. Gascón, S. et al. Identification and successful negotiation of a metabolic checkpoint in direct neuronal reprogramming. *Cell Stem Cell* **18**, 396–409 (2016).
This work revealed a metabolic hurdle during the direct reprogramming of glia and many other cell types into neurons and showed that protecting neurons from death and ROS resulted in very high conversion efficiency in the injured mouse cerebral cortex in vivo.
85. Wang, L. L. & Zhang, C. L. Engineering new neurons: in vivo reprogramming in mammalian brain and spinal cord. *Cell Tissue Res.* **371**, 201–212 (2018).
86. Southwell, D. G., Froemke, R. C., Alvarez-Buylla, A., Stryker, M. P. & Gandhi, S. P. Cortical plasticity induced by inhibitory neuron transplantation. *Science* **327**, 1145–1148 (2010).
87. Tang, Y., Stryker, M. P., Alvarez-Buylla, A. & Espinosa, J. S. Cortical plasticity induced by transplantation of embryonic somatostatin or parvalbumin interneurons. *Proc. Natl Acad. Sci. USA* **111**, 18339–18344 (2014).
88. Dimou, L. & Götz, M. Glial cells as progenitors and stem cells: new roles in the healthy and diseased brain. *Physiol. Rev.* **94**, 709–737 (2014).
89. Heinrich, C. et al. Sox2-mediated conversion of NG2 glia into induced neurons in the injured adult cerebral cortex. *Stem Cell Reports* **3**, 1000–1014 (2014).
90. Niu, W. et al. SOX2 reprograms resident astrocytes into neural progenitors in the adult brain. *Stem Cell Reports* **4**, 780–794 (2015).
91. Liu, Y. et al. Ascl1 converts dorsal midbrain astrocytes into functional neurons in vivo. *J. Neurosci.* **35**, 9336–9355 (2015).
92. Deverman, B. E. et al. Cre-dependent selection yields AAV variants for widespread gene transfer to the adult brain. *Nat. Biotechnol.* **34**, 204–209 (2016).
93. Vagner, T., Dvorzhak, A., Wójtowicz, A. M., Harms, C. & Grantyn, R. Systemic application of AAV vectors targeting GFAP-expressing astrocytes in Z-Q175-KI Huntington's disease mice. *Mol. Cell. Neurosci.* **77**, 76–86 (2016).
94. Kunze, C. et al. Synthetic AAV/CRISPR vectors for blocking HIV-1 expression in persistently infected astrocytes. *Glia* **66**, 413–427 (2018).
95. Pereira, M. et al. Direct reprogramming of resident NG2 glia into neurons with properties of fast-spiking parvalbumin-containing interneurons. *Stem Cell Reports* **9**, 742–751 (2017).
This work shows how a very precise neuronal subtype can be achieved by targeting a specific glial subtype.
96. Rivetti di Val Cervo, P. et al. Induction of functional dopamine neurons from human astrocytes in vitro and mouse astrocytes in a Parkinson's disease model. *Nat. Biotechnol.* **35**, 444–452 (2017).
97. Karow, M. et al. Reprogramming of pericyte-derived cells of the adult human brain into induced neuronal cells. *Cell Stem Cell* **11**, 471–476 (2012).
This study showed that non-neuronal cells from the adult human brain could be successfully converted into functional neurons, opening up the possibility of using direct in vivo neuronal reprogramming for human patients.
98. Pfisterer, U. et al. Direct conversion of human fibroblasts to dopaminergic neurons. *Proc. Natl Acad. Sci. USA* **108**, 10343–10348 (2011).
99. Drouin-Ouellet, J. et al. REST suppression mediates neural conversion of adult human fibroblasts via microRNA-dependent and -independent pathways. *EMBO Mol. Med.* **9**, 1117–1131 (2017).
100. Pang, Z. P. et al. Induction of human neuronal cells by defined transcription factors. *Nature* **476**, 220–223 (2011).
101. Zhang, L. et al. Small molecules efficiently reprogram human astroglial cells into functional neurons. *Cell Stem Cell* **17**, 735–747 (2015).

Acknowledgements R.A.B. is funded by the NIHR Biomedical Research Centre in Cambridge, Cure PD, PDUK, European Research Council under the European Union's Seventh Framework Programme: FP/2007-2013 NeuroStemcellRepair (no. 602278), Wellcome Trust MRC Stem Cell Institute and MRC UKRMP PSCP. He has received consultancy payments from FCDI and LCT. M.G. is funded by the German Research Foundation (CRC870, SPP1738, 1757, EXC1010 Synergy), The Ministry of Science and Research (MAIV), ERANET and the ERC (ChroNeuroRepair). M.P. receives funding from the New York Stem Cell Foundation, the European Research Council under the European Union's Seventh Framework Programme: FP/2007-2013 NeuroStemcellRepair (no. 602278) and ERC Grant Agreement no. 30971, the Swedish Research Council and the Strategic Research Area Multipark at Lund University. M.P. is a New York Stem Cell Foundation Robertson Investigator. We thank D. Daft for her help in the preparation of this manuscript.

Reviewer information *Nature* thanks P. Arlotta, J. Takahashi and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions All three authors contributed to the design, writing and critical review of this manuscript.

Competing interests M.P. is the owner of Parmar Cells AB and co-inventor on US patent applications 15/093,927 owned by Biolamina AB and EP17181588 owned by Miltenyi Biotec. Patent WO 2015/114059 A1 patents the use of BCL2 in reprogramming. R.A.B. and M.G. have no competing interests.

Additional information
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to R.A.B. or M.G. or M.P.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

https://doi.org/10.1038/s41586-018-0089-z

Bioengineering strategies to accelerate stem cell therapeutics

Christopher M. Madl[1], Sarah C. Heilshorn[2] & Helen M. Blau[1]*

Although only a few stem cell-based therapies are currently available to patients, stem cells hold tremendous regenerative potential, and several exciting clinical applications are on the horizon. Biomaterials with tuneable mechanical and biochemical properties can preserve stem cell function in culture, enhance survival of transplanted cells and guide tissue regeneration. Rapid progress with three-dimensional hydrogel culture platforms provides the opportunity to grow patient-specific organoids, and has led to the discovery of drugs that stimulate endogenous tissue-specific stem cells and enabled screens for drugs to treat disease. Therefore, bioengineering technologies are poised to overcome current bottlenecks and revolutionize the field of regenerative medicine.

Stem cell therapies have the potential to transform medicine by enabling patient-specific regeneration of injured or diseased tissues, providing cures for some of humanity's most intractable diseases, such as muscular dystrophies, diabetes and neurodegeneration. The rapid expansion of stem cell research over the past two decades has uncovered methods that use a patient's own cells to form mature cell types and even miniature organs, or organoids, in the laboratory. These strategies can harness the native regenerative capacity of somatic stem cells that reside in the patient's own tissues, such as the bone marrow or skeletal muscle. Alternatively, the advent of induced pluripotent stem (iPS) cells allows researchers to take mature cells from a patient's skin or blood and reprogram these cells into an immature, embryonic state. These iPS cells can then be differentiated into any cell type of any given adult tissue, providing an avenue to achieve the goals of personalized medicine. Such patient-specific cells can be used to repair damaged tissues or as diagnostic tools to screen for drugs or inform treatment decisions made by physicians.

Successful reports of translating stem cell therapies to patients over the past several years have fostered hope that strategies for regenerative medicine may one day cure some of the most challenging illnesses. Recently, genetically modified keratinocyte cultures containing epidermal stem cells restored more than 80% of the surface area of the skin of a young patient suffering from a deadly blistering disorder[1]. In other examples, embryonic stem (ES) cells or patient-derived iPS cells that were differentiated into retinal pigment epithelial cells and transplanted into the eye improved the sight of patients at risk of becoming blind due to macular degeneration[2–4]. Despite such highly publicized and exciting cases of success, the majority of stem cell clinical trials to date have not yet achieved regulatory approval and commercialization as stem cell therapies[5]. Although hundreds of clinical trials are registered with the US FDA (Food and Drug Administration) on the clinical trials website (https://clinicaltrials.gov/), the only FDA-approved stem cell products consist of umbilical cord blood-derived haematopoietic progenitors[6]. World-wide, rigorous clinical trials have led to approval of only a handful of therapies based on adult stem cells[5]. This is not only because of lengthy regulatory hurdles, but also due to biological obstacles.

Despite substantial advances in our understanding of stem cell biology, several challenges remain that limit the widespread clinical use of stem cell therapies. Current hurdles to the clinical translation of stem cell therapies include maintenance of the stem cell state, reproducible expansion of large numbers of stem cells for transplantation, efficient control of the cell state both pre- and post-transplantation, and protection of the cells during and after delivery to patients (Fig. 1). Another bottleneck is exemplified by a failure of clinical-grade neural stem cells to replicate the regenerative effects of research-grade cells in pre-clinical animal models, highlighting the difficulties associated with stem cell production and transplantation for use in patients[7,8]. Engineering approaches offer solutions to overcome current limitations. In particular, advances in materials science have enabled unprecedented control over the biochemical and biophysical properties of materials used for stem cell therapies. Material

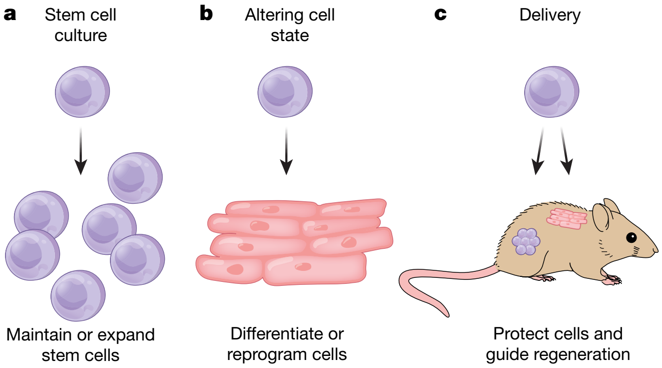


Fig. 1 | Challenges in translating stem cell therapies with potential bioengineered solutions. a, Present challenges culturing stem cells include maintenance of the stem cell state ex vivo[16,18,20,21,23–25] and efficient expansion of naive stem cells[9,10,12,27]. **b**, To fully realize the potential of stem cells, reliable protocols for altering cell state must be developed, including differentiation of stem cells to mature cell types[35–45] and reprogramming of somatic cells to pluripotent stem cells[47,48]. **c**, Conventional cell delivery approaches do not address crucial obstacles in cell transplantation therapies, including maintaining the viability and potency of stem cells during injection[49–51], providing a supportive microenvironment for the cells after implantation[50,52–55], and controlling the fate of the cells by providing cues to guide regeneration in vivo[56,59]. Engineering approaches are being applied to design materials to address these challenges.

---

[1]Baxter Laboratory for Stem Cell Biology, Department of Microbiology and Immunology, Stanford University School of Medicine, Stanford, California, USA. [2]Department of Materials Science and Engineering, Stanford University, Stanford, California, USA. *e-mail: hblau@stanford.edu

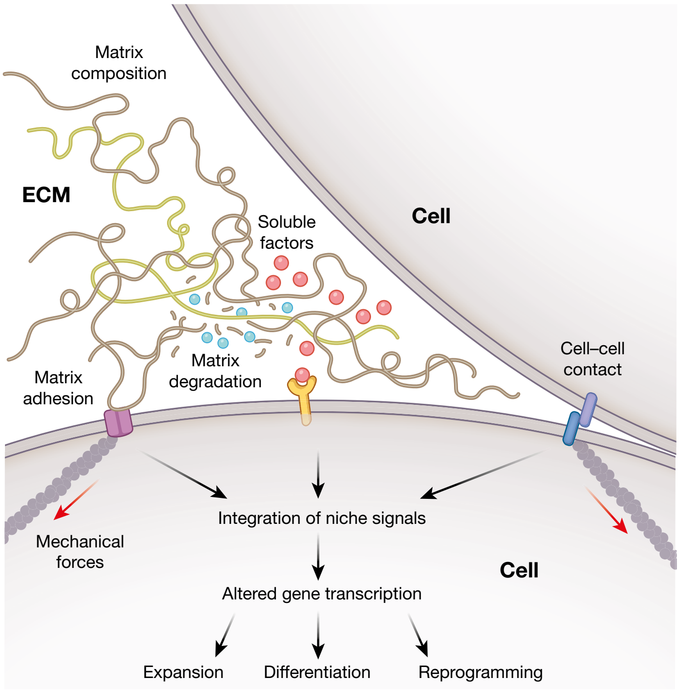


Fig. 2 | Recapitulating niche interactions to direct stem cell fate. Various biochemical and biophysical factors within the stem cell microenvironment combine to modulate cellular behaviours. Careful design of materials for stem cell culture and transplantation can effectively control matrix properties, such as biochemical composition, mechanics and degradation, as well as soluble factor signalling and cell–cell contact to regulate stem cell fate.

properties can be tuned to create an artificial niche to both expand naive stem cells and efficiently differentiate stem cells into mature cell types (Fig. 2). Material carriers can improve the survival and engraftment of transplanted stem cells, and controlling the properties of these carriers can promote an enhanced regenerative response from the delivered cells. Innovative material design can aid in meeting regulatory standards and facilitate the increase in scale necessary for commercialization. Here we describe how bioengineered materials have already substantially contributed to stem cell advances and discuss how novel material design can overcome remaining difficulties, accelerating and expanding clinical applications of stem cell-based therapies.

Expanding stem cells

One of the major bottlenecks in translating stem cell therapies to the clinic has been expansion of the large numbers of cells that are required for transplantation, typically tens to hundreds of millions of cells per patient. Cells to be used for human therapies must be cultured under fully defined conditions to meet regulatory requirements and exhibit minimal batch-to-batch variation for consistent therapeutic efficacy. Furthermore, the platforms used for stem cell expansion must be amenable to industrial scale-up.

Historically, the most common techniques used to culture pluripotent stem cells (ES and iPS cells) contained animal-derived components, such as a layer of live mouse embryonic fibroblasts or coatings of Matrigel, a mixture of extracellular matrix (ECM) proteins derived from mice. The use of such animal-derived components hinders regulatory approval, and ECM extracts such as Matrigel can be highly variable in their composition, potentially impacting the reproducibility of cells expanded by using this material. To address these issues, coatings for tissue culture substrates have been designed to recapitulate specific aspects of the native biochemistry. Fully recombinant ECM proteins[9], surfaces grafted with peptides that promote cell adhesion[10,11], and synthetic polymer coatings[12] have all been used to facilitate expansion of pluripotent stem cells. Additionally, high-throughput techniques have enabled screening for matrix-bound protein components and their effects on stem cell state, further refining the presentation of microenvironmental cues from biomimetic materials. For instance, micro-well arrays spotted with proteins led to identification of niche factors that promoted neural stem cell (NSC) proliferation[13].

In addition to the biochemical composition, the physical properties of the matrix on which stem cells reside can also alter stem cell expansion. One crucial physical regulator of stem cell fate is matrix stiffness, a measure of how easily the matrix deforms under an applied load. The stiffness of biological materials is often reported as an elastic modulus (also known as the Young's modulus), which is an inherent material property, independent of material geometry. In the body, stem cells and their progeny experience stiffness spanning several orders of magnitude, from relatively compliant brain tissue (elastic moduli of approximately 10^2 Pa)[14] to rigid calcified bone (elastic moduli of approximately 10^{10} Pa)[15]. Systems with a tunable stiffness that encompasses a physiological range have typically used materials known as hydrogels, which are water-swollen polymer networks.

The profound role of substrate stiffness in regulating the self-renewal of somatic stem cells is clear from studies using hydrogel substrates. For example, muscle stem cells (MuSCs), which are responsible for the maintenance and repair of skeletal muscle tissue, need to be cultured on hydrogels with an equivalent stiffness to native muscle tissue to maintain their regenerative potential during expansion in culture and after transplantation in vivo[16]. Moreover, substrate elasticity has had a crucial role in the ability to 'rejuvenate' MuSCs derived from aged mice to improve regenerative function[17]. With ageing, MuSCs acquire intrinsic defects that make them less potent than cells derived from young mice[17], hampering the much-needed therapeutic function of native MuSCs in elderly individuals. However, a combination of culturing aged MuSCs on compliant substrates with muscle-like stiffness and pharmacological inhibition of p38 MAP kinase resulted in expansion of a stem cell pool with improved engraftment and regenerative capacity, culminating in a marked increase in strength[17].

The observation that substrate stiffness in culture can regulate the function of expanded stem cells even after transplantation in vivo suggests that the stem cells are capable of 'remembering' the mechanical environment in which they were cultured. A noteworthy study, which used hydrogels that dynamically soften in response to controlled light exposure, has shown that mesenchymal stem cells (MSCs), which are bone, cartilage and fat-forming cells derived from the bone marrow, possessed a 'mechanical memory'[18]. Stiff hydrogels biased the MSCs towards differentiation over stem cell maintenance, and prolonged culture on stiff substrates resulted in an irreversible loss of stem cell potential[18]. Identification of the molecular mechanisms that are responsible for this mechanical memory may help to restore function in stem cells that have acquired defects from fibrotic stiffening due to disease and ageing. One such memory molecule in MSCs is the microRNA miR-21. Resetting expression levels of miR-21 effectively 'erased' the memory of being cultured on a stiff substrate[19].

Other stem cell types are similarly sensitive to stiffness and expand optimally when cultured on substrates of a particular elasticity. Culture on compliant substrates enhanced the ex vivo expansion of haematopoietic stem cells[20,21], which are responsible for reconstituting blood and immune cells. Human embryonic stem cells were best maintained on relatively compliant substrates, which led to the expression of high levels of pluripotency genes and retention of the capacity of these cells to differentiate into all three germ layers[22].

The two-dimensional (2D) nature of traditional cell culture often does not adequately replicate the three-dimensional (3D) environment experienced by stem cells in the body. Hydrogels have proven to be a useful material platform to culture cells in a more native-like 3D microenvironment. Studies of the native ECM revealed that cell-secreted enzymes, such as metalloproteases, remodel the matrix to permit cell spreading and migration through their surrounding material. Incorporating this principle, hydrogel systems for 3D culture have been engineered to permit degradation and remodelling. Matrix remodelling has recently been demonstrated to have significant and diverse impacts on the expansion of stem cells ex vivo. NSCs embedded within 3D hydrogels must remodel the surrounding matrix in order to maintain cell–cell contacts and retain their stem cell state, irrespective of matrix stiffness[23]. By contrast, maintenance and proliferation of intestinal stem cell cultures is decreased upon culture in hydrogels susceptible to degradation by cell-secreted enzymes[24]. However, to facilitate maturation of intestinal organoid cultures, gradual, passive degradation of the matrix is necessary[24].

Microstructural variation is an additional parameter provided by the native ECM that is not replicated in traditional 2D cultures or homogeneous 3D hydrogels. The ECM of many tissues consists of fibrous components spanning the nano- to micrometre scales. Cell culture substrates presenting features along these length scales can alter cellular behaviour. In one example, nanoscale-patterned surfaces with a square lattice geometry promoted enhanced maintenance of a stem cell phenotype in cultured MSCs[25]. ES cells are also acutely sensitive to nanoscale topography. Culture on nanoscale smooth surfaces promoted ES cell self-renewal and maintenance of pluripotency, whereas culture on nanoscale rough surfaces induced spontaneous differentiation[26].

Engineered materials can address processing concerns related to industrial scale-up of stem cell production. Traditional 2D culture methods have high space and nutrient costs. Transitioning to 3D culture platforms can decrease the amount of surface area required for cell culture by stacking cells in the *z*-dimension. To this end, temperature-responsive 3D hydrogel systems have been developed for easy encapsulation and expansion of pluripotent stem cells[27]. In addition to saving space, these hydrogels facilitate the collection of the expanded stem cells, which can be triggered by simply lowering the temperature to dissolve the polymers comprising the gel[27]. Successful commercialization of stem cell therapies will ultimately require large-scale cell culture technologies, such as bioreactors[28,29]. Many stem cells need to adhere to surfaces to maintain their stem cell state, requiring materials that can serve as microcarriers that provide both crucial chemical and mechanical cues to cells cultured in large reactors. Good examples of these microcarriers are polymeric microbeads coated with matrix proteins as supports in stirred reactors[30,31] and hydrogel microbeads that facilitate adhesion and expansion of pluripotent stem cells in reactor-compatible 3D microenvironments[32].

Altering cell state

Many of the proposed therapeutic applications of stem cells require controlled methods of altering cell state. For tissue-replacement therapies using stem cells differentiated into mature cell types, highly efficient differentiation into the target cell population is required to limit potential deleterious effects of co-transplanting either highly proliferative naive stem cells or other potentially antagonistic differentiated cells. This concern also applies to in vitro studies of cells derived from stem cells, as the presence of improperly differentiated cells can skew the results of bulk biochemical assays. Just as physical matrix properties can be used to preserve stem cell phenotype, these matrix properties can be tuned to direct and augment the differentiation of stem cells.

The first demonstration that physical interactions with the matrix could mediate changes in cell state arose from seminal studies of malignant transformation in breast cancer. In 3D cultures of malignant breast cancer cells, reversion of the cells to a non-malignant phenotype was achieved by blocking specific integrins, the cell-surface receptors that connect the intracellular force-generation mechanisms of the cytoskeleton to the ECM[33]. Conversely, increasing matrix stiffness resulted in transformation of cells from a benign to a malignant phenotype[34]. These studies pointed to force generation by cells as a means for sensing and responding to the mechanical properties of the matrix.

In a ground-breaking study, the differentiation of MSCs was demonstrated to be biased according to the stiffness of their underlying substrate[35]. MSCs cultured on stiff substrates similar to pre-calcified bone preferentially differentiated into bone cells, whereas MSCs cultured on intermediate stiffness similar to muscle tissue displayed a more muscle-like phenotype[35]. MSCs cultured on the most compliant matrices, reminiscent of brain tissue, exhibited a neuron-like phenotype[35].

The premise that matrix stiffness can alter cell state through force generation was corroborated in 3D materials, with an optimal stiffness mediating bone differentiation of MSCs through integrin clustering[36]. NSCs are also sensitive to matrix mechanics, preferentially differentiating into neurons on very compliant substrates similar to the elasticity of brain tissue and into supporting glial cells on stiff substrates[37]. Furthermore, matrix stiffness may have a crucial role in developmental processes, as compliant substrates have been shown to enhance the mesodermal differentiation potential of embryonic stem cells[38]. Beyond materials with a fixed stiffness, recent studies have implicated the time-dependent mechanical properties of viscoelastic materials as regulators of MSC differentiation[39–41] (Box 1). At a fixed stiffness, MSC differentiation into a bone lineage was markedly enhanced in materials with a greater viscous character[40].

Various other matrix parameters have been implicated in regulating how stem cells alter their cell state. For instance, matrix degradation by encapsulated MSCs was required for force generation and subsequent differentiation into a bone lineage[42]. Cell–cell contacts can alter how MSCs respond to mechanical cues[43]. The cell-adhesive ligands presented by the matrix also play an important part in directing differentiation. By controlling the temporal presentation of adhesive cues, MSC differentiation into cartilage[44] and NSC differentiation into neurons[45] were enhanced. Because the interactions among multiple types of cell-adhesive ligands are often complicated and nonlinear, combinatorial studies that include statistical approaches have been used to optimize ligand composition to promote differentiation[46].

The reprogramming of somatic cells to generate patient-specific iPS cells is also highly sensitive to matrix interactions that modulate cell state. Traditional iPS cell reprogramming protocols use standard 2D tissue culture techniques. However, the properties of the culture substrate can have a substantial impact on the efficiency of iPS cell colony generation. Forcing alignment of fibroblasts on substrates with aligned microgrooves mediated epigenetic modifications that increased reprogramming efficiency[47]. Transitioning to a 3D hydrogel platform also resulted in an increase in efficiency, with optimal matrix properties identified via high-throughput screens[48]. Such technologies have the potential to decrease the variability and cost associated with generating patient-specific stem cell therapies.

Improving cell delivery

Efficient transplantation and engraftment into host tissues remains a notable barrier to therapeutic success. Many cells die from the mechanical damage that is caused by the injection process or fail to engraft in the relatively inhospitable microenvironment of damaged tissue. Recent advances in material-based cell delivery systems show promise in overcoming these difficulties.

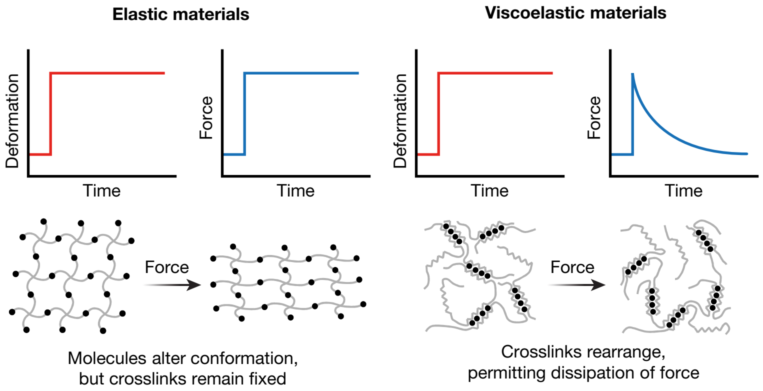
The simple act of injecting stem cells through a needle significantly reduces the viability of the injected cells[49,50]. As the solution in which the cells are suspended transitions from the syringe barrel to the needle, the fluid undergoes an increase in velocity of two orders of magnitude, exposing the cells to substantial extensional flows that can damage cell membranes[49]. The porous and highly hydrated nature of hydrogels is ideally suited to encapsulation of small molecules, growth factors or proteins together with stem cells. Injectable hydrogels have been developed to limit the membrane damage experienced by cells[49,51]. In these systems, the bulk of the hydrogel moves through the needle as a solid; only the edges near the needle wall flow like a liquid[51]. Therefore, the vast majority of the cells pass through the needle without experiencing damaging shear deformation through extensional flow.

Once injected into tissue, hydrogel carriers can also serve to retain the cells at the target location. Very few cells that are delivered via commonly used saline injections are retained in tissues for an extended period. Rather, the immune response of the host in the damaged tissue and the lack of adhesive sites lead to cell death and clearance. Accordingly, by increasing the stability of injected hydrogel carriers,

BOX 1

Elasticity versus viscoelasticity

The term stiffness is commonly used to describe materials that exhibit elastic mechanical properties. When a force is applied to an elastic material, the force is retained in the material over time, similar to the way in which a strained rubber band provides constant resistance (see figure). The native ECM is not a purely elastic material, but rather exhibits both elastic (solid-like) and viscous (liquid-like) behaviour. Viscoelastic materials have time-dependent mechanical properties. For instance, natural ECM exhibits stress-relaxation, where resistance to an applied load is dissipated over time by rearrangement of the molecules that comprise the ECM (see Figure). This is analogous to a ball of putty deforming over time after a force is applied. Thus, characterizing materials used in cell culture by only stiffness may oversimplify the mechanisms by which cells can interact with bioengineered materials. Recent studies have highlighted the importance of accounting for viscoelasticity in stem cell differentiation. At a given stiffness, both in 2D and 3D, tuning the viscous characteristics of the material can enhance the differentiation of MSCs into bone[39–41].



cell retention was increased at the injection site in a mouse model[50]. Moving towards therapeutic applications, injectable hydrogel-mediated delivery of endothelial progenitor cells to ischaemic rat hearts increased cell engraftment and decreased fibrosis compared to cells injected in saline[52]. Injectable hydrogels have also significantly improved the survival of iPS cell-derived oligodendrocytes delivered to injured rat spinal cords[53].

Hydrogels can serve as immuno-protective barriers to shield the transplanted cells from host inflammation, overcoming a major difficulty for the use of allogeneic cells for transplantation. A particularly poignant case in point is the treatment of diabetes by pancreatic cells that comprise the β-islets. The ideal material would protect the transplanted cells from immune clearance while permitting sustained insulin secretion. Early approaches met with limited clinical success in part because of immune responses to both animal-derived cells and the materials intended to protect these cells. Islets derived from human ES cells and hydrogel materials that elicit minimal inflammatory responses have provided new hope that islet transplantation can be used to cure type I diabetes[54]. The immuno-protective effect of hydrogels may also facilitate cell-mediated tissue regeneration, as hydrogel delivery of iPS cell-derived neural progenitors has been shown to decrease inflammation and improve neuronal differentiation compared to saline delivery[55].

To enhance the regenerative phenotypes of the delivered cells, some of the same matrix properties used to modulate cell state in culture can be incorporated into hydrogel delivery vehicles. For instance, differentiation of MSCs towards a bone lineage is known to be mechanosensitive in vitro[35,36], and transplanting MSCs in hydrogels of optimal stiffness enhanced bone regeneration in a critical-sized cranial defect model in vivo[56]. Furthermore, the recent observation that MSC differentiation in vitro depends not only on the time-independent elasticity of the material, but also on the time-dependent viscoelasticity of the material[39–41] (Box 1), also holds true when viscoelastic hydrogels are used to transplant MSCs in vivo. Cells delivered in hydrogels with a more viscous character exhibited increased bone regeneration in vivo compared to hydrogels of a comparable stiffness that were predominantly elastic[57].

Hydrogel microstructure can also profoundly impact the fate of transplanted stem cells. Peptide amphiphile hydrogels are a classic example of a 3D cell culture material with a characteristic nanoscale fibrous architecture[58]. When these materials are subjected to heating and cooling, the peptides self-assemble to form noodles of aligned nanofibres that, when mixed in a calcium-rich suspension, can encapsulate cells together with growth factors[59]. These scaffolds increase cell viability, mediate cell alignment parallel to the hydrogel nanofibres, and have degradation rates that fit the time course of regeneration[59]. These features are ideally suited to MuSC delivery and have led to improved MuSC engraftment and muscle repair[59].

Successful regeneration of functional tissue requires integration of transplanted stem cells with the host vasculature and innervation of the newly formed tissue. Neovascularization is critical for long-term survival of transplanted cells, as oxygen and nutrient transport requirements dictate that, in general, cells must be located within 100–200 μm of a capillary[60]. Classical strategies for vascularization have taken a bottom-up approach starting from the individual cellular components of blood vessels, relying on self-assembly of either host or exogenous endothelial cells and smooth muscle cells. Sequential delivery of angiogenic factors from biomaterial scaffolds resulted in initial vascular sprouting followed by vessel maturation culminating in a more robust vasculature[61]. Alternative approaches have used co-cultures of endothelial cells, supporting stromal cells and tissue progenitor cells to generate vascularized tissues that were perfused by host vasculature when transplanted[62]. More recently, top-down techniques starting from the viewpoint of the finalized tissue, including 3D printing[63] and two-photon lithography[64], have been used to produce engineered constructs with user-defined vasculature. Innervation of engineered tissues can be achieved through delivery of growth factors, including classical neurotrophins such as nerve growth factor[65] and angiogenic factors such as vascular endothelial growth factor[66].

Improving human cell culture models

Cell transplantation for tissue regeneration is just one facet of personalized medicine made possible by advances in stem cell biology. Patient-derived iPS cells that can give rise to a myriad of differentiated cell types have provided researchers unprecedented access to diverse healthy and diseased samples that can help to inform basic biology, drug screens and toxicology studies. A challenge has been the degree of differentiation, as the differentiated cell types obtained are notoriously immature. Engineered cellular microenvironments offer hope.

Engineered matrices have been developed to support the complex 3D architecture of organotypic cultures that are used to study developmental and disease processes. Synthetic hydrogels can replace the highly variable Matrigel substrate in primary[24,67] as well as ES and iPS cell-derived[68] intestinal organoids. Microfilament scaffolds improved cortical development in human brain organoids[69], and controlling matrix stiffness and 2D versus 3D dimensionality permitted generation of amnion-like structures[70].

Engineered human tissue constructs provide a novel platform to study disease progression and test potential therapeutic interventions. In a noteworthy example, human iPS cell-derived brain organoids produced by bioreactor culture and infected with Zika virus exhibited reduced NSC proliferation, suggestive of a microcephaly-like phenotype[71]. Engineered tissues also enable the study of patient-specific genetic diseases. Filamentous matrices were used to generate cardiomyocytes from healthy and diseased patients to study contractile abnormalities in congenital cardiomyopathy[72].

The patient-specific nature of iPS cell-derived cells makes them attractive platforms for screening drugs for potential toxicity on an individual patient level. For example, cardiomyocytes generated from patient-derived iPS cells recapitulate the heightened toxicity in response to chemotherapy seen in specific cancer patients[73]. Advances in engineered microsystems have provided platforms to investigate the effects of drug treatment on iPS cell-derived cardiomyocyte function[74,75]. In addition to cardiac models, synthetic matrices have been used to improve the sensitivity of vascular toxicity screens[76], and microphysiological systems using iPS cell-derived kidney cells recapitulate drug-induced kidney toxicity[77].

Future outlook

Although only a handful of stem cell therapies have currently been approved for use in patients, several exciting clinical applications are on the horizon that have benefitted from bioengineered materials (Fig. 3). Rapid progress is being made in the use of organotypic cultures from patient-derived stem cells or tissue-specific stem cells in hydrogels[24,67,68,70,78,79]. This advance has led to the discovery of drugs to treat disease. A striking example highlights how patient-specific organoid cultures can profoundly impact clinical outcome. Intestinal organoids derived from patients with cystic fibrosis were grown in 3D hydrogels and used to screen drugs that could reverse the effects of the disease[78]. These culture models have uncovered life-changing therapies for patients suffering from very rare mutations, or orphan diseases, by rapidly and effectively assessing potential efficacy of costly drugs[80]. As a result, patients with cystic fibrosis who once had no treatment options have now been matched with drugs that address their disease in culture, ameliorate their symptoms, and markedly improve their quality of life[80]. In another example, 3D organoid cultures of cochlear stem cells enabled identification of a combination of small molecules that can stimulate expansion of these cells, which in turn differentiate into hair cells responsible for hearing[79]. This drug combination may enable the activation of endogenous stem cells to reverse hearing loss in patients[81], a problem that confronts our increasingly aged population. Combining advances in hydrogel stem cell culture techniques with in silico screens can further increase the success rate of identifying new drugs targeting endogenous stem cells. Such an in silico approach was instrumental in identifying prostaglandin E2 as a natural inflammatory modulator capable of potently inducing of skeletal muscle stem cell expansion in vitro and muscle regeneration in vivo[82]. This approach capitalizes on the quiescent stem cells resident in muscle tissues throughout life that are dedicated to skeletal muscle repair. The function of these cells declines with age[17]. Identification of agents capable of rejuvenating the function of these endogenous stem cells opens the door to therapies that counter muscle wasting and restore strength, countering frailty, a major cause of morbidity with ageing.

Fully realizing the potential of personalized medicine provided by stem cells will require advances in bioengineered materials. The native stem cell microenvironment is highly dynamic, with temporally varying biochemical and biophysical properties. For example, tissue dysfunction in ageing and disease is often characterized by fibrosis, which reflects

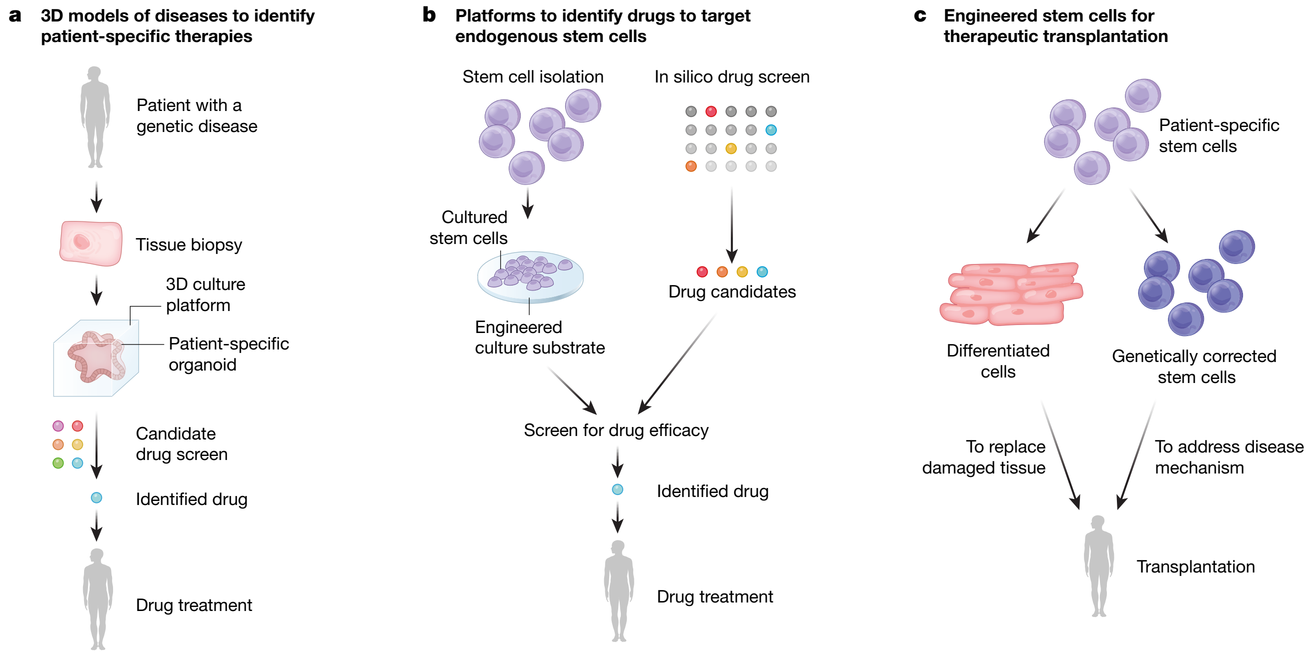


Fig. 3 | Impact of bioengineering on stem cell advances currently in the clinic or on the horizon. a, **b**, Hydrogel-based culture systems, such as intestinal organoid cultures, have enabled identification of promising drugs to treat cystic fibrosis[78,80] (**a**), while others are used to target endogenous stem cells within tissues to restore hearing[79,81] and augment strength[17,82] (**b**). **c**, Treatments in clinical trials that could achieve greater efficacy by using engineered scaffolds to culture and transplant cells include ES and iPS cell-derived retinal epithelial cells to restore vision to macular degeneration patients[2–4] and skin grafts of genetically corrected epidermal stem cells to save patients from a deadly skin blistering disease[1].

an increase in deposition and crosslinking of ECM proteins that leads to changes in the stiffness and composition of the cellular microenvironment. Given that cells are acutely sensitive to these signals, biomimetic materials are needed that permit incorporation of this dynamism to enable improved in vitro models of fibrosis. Recent studies have shown promise using light- and enzyme-mediated approaches to alter substrate stiffness and presentation of bioactive factors[18,44,83–87]. In particular, photo-mediated degradation of hydrogel crosslinks has enabled dynamic softening of cell culture substrates to study how stem cells respond to changes in their mechanical environment[18]. Conversely, photo-initiated polymerization has enabled in situ stiffening of hydrogels[87], reminiscent of fibrotic disease states. Combining these approaches with genetic reporters commonly used in cell biology settings may enable real-time investigation of signalling changes as a result of changes in matrix properties. However, existing chemical approaches to dynamically modulate matrix mechanics are commonly limited by the use of potentially mutagenic UV light, although chemistries that are compatible with two-photon and blue-light illumination have been developed[44,83]. Furthermore, the free radicals generated during photo-initiated polymerization and stiffening may be toxic to sensitive stem cells. Future work directed at refining these approaches should focus on using less perturbative stimuli, such as visible light and exploring fully biocompatible chemistries, to permit completely orthogonal tuning of the cellular microenvironment both in vitro and in vivo. Such dynamically tunable systems may improve the accuracy of preclinical models by better mimicking the native cellular niche and potentially enable modulation of tissue-engineered constructs in vivo to facilitate precise spatiotemporal control of morphogenic cues.

Furthermore, we must increase our understanding of the critical role of endogenous tissue-specific cell modifications of delivered materials. In addition to potential effects on endogenous stem cells, impacts on immune cells, which are crucial to efficacious regeneration, are of paramount importance. Many existing strategies for material design have focused on minimizing the immune response of the host at the site of delivery. This is particularly important for the transplantation of allogeneic, as opposed to patient-specific, stem cell therapies. Initial widespread application of iPS cell therapies may utilize libraries of reprogrammed stem cell lines to enable close genetic matching between donor and recipient, similar to the way that organ donations are screened for a close antigenic match[88]. Such allogeneic approaches do not guarantee successful transplant engraftment and may still require immunosuppression, but may provide earlier access to stem cell therapies until systems for generating and validating patient-specific stem cells are implemented on a large scale. Lessons learned from the development of biomaterials for immuno-isolated tissue engineering, for instance in islet transplantation for diabetes treatment[54], can be applied to generate materials that enhance stem cell engraftment and function by limiting deleterious local immune responses.

The direct participation of immune cells in the process of regeneration is increasingly recognized as essential to proper restoration of tissue function. Thus, in contrast to materials designed to evade an immune response, well-designed immunomodulatory materials could aid in the process of regeneration[89,90]. Although studies to date have mostly focused on using materials to direct the immune response to existing disease states, such as targeting cancer[91,92] or induction of tolerance in autoimmune disorders[93,94], lessons learned from programming immune cells with biomaterials can be harnessed to improve stem cell therapeutic outcomes by orchestrating the immune response during regeneration. For instance, self-assembling peptide scaffolds have been used to modulate presentation of T cell epitopes, resulting in a dose-dependent response to activate different immune cell populations[95]. Similar material strategies may be used in tandem with stem cell-targeting factors to simultaneously regulate the immunological response during tissue regeneration. Initial results using materials to control cytokine delivery enabled temporal control over recruitment of different macrophage subtypes, which in turn secreted different angiogenic factors at appropriate morphogenic time points to enhance vascularization of tissue engineered constructs[96].

As novel materials are developed for stem cell therapies, regulatory requirements must also be considered. Competing interests in designing materials that sufficiently recapitulate the complexities of the native matrix to control cell fate must be balanced with the need to develop scalable, cost-effective platforms for commercialization. For clinical use, materials must be fully defined and free of animal-derived components. By synthesizing knowledge from fields such as materials science, chemistry, bioengineering and cell biology, the groundwork has been laid to propel stem cell-based therapies into the clinic.

Received: 15 January 2018; Accepted: 16 March 2018;
Published online 16 May 2018.

1. Hirsch, T. et al. Regeneration of the entire human epidermis using transgenic stem cells. *Nature* **551**, 327–332 (2017).
2. Schwartz, S. D. et al. Embryonic stem cell trials for macular degeneration: a preliminary report. *Lancet* **379**, 713–720 (2012).
3. Schwartz, S. D. et al. Human embryonic stem cell-derived retinal pigment epithelium in patients with age-related macular degeneration and Stargardt's macular dystrophy: follow-up of two open-label phase 1/2 studies. *Lancet* **385**, 509–516 (2015).
4. Mandai, M. et al. Autologous induced stem-cell-derived retinal cells for macular degeneration. *N. Engl. J. Med.* **376**, 1038–1046 (2017).
5. Trounson, A. & McDonald, C. Stem cell therapies in clinical trials: progress and challenges. *Cell Stem Cell* **17**, 11–22 (2015).
6. FDA warns about stem cell therapies. *US Food & Drug Administration* https://www.fda.gov/ForConsumers/ConsumerUpdates/ucm286155.htm (FDA, 2017).
7. Anderson, A. J., Piltti, K. M., Hooshmand, M. J., Nishi, R. A. & Cummings, B. J. Preclinical efficacy failure of human CNS-derived stem cells for use in the pathway study of cervical spinal cord injury. *Stem Cell Reports* **8**, 249–263 (2017).
8. Marsh, S. E. et al. HuCNS-SC Human NSCs fail to differentiate, form ectopic clusters, and provide no cognitive benefits in a transgenic model of Alzheimer's disease. *Stem Cell Reports* **8**, 235–248 (2017).
9. Rodin, S. et al. Long-term self-renewal of human pluripotent stem cells on human recombinant laminin-511. *Nat. Biotechnol.* **28**, 611–615 (2010).
10. Melkoumian, Z. et al. Synthetic peptide-acrylate surfaces for long-term self-renewal and cardiomyocyte differentiation of human embryonic stem cells. *Nat. Biotechnol.* **28**, 606–610 (2010).
11. Klim, J. R., Li, L., Wrighton, P. J., Piekarczyk, M. S. & Kiessling, L. L. A defined glycosaminoglycan-binding substratum for human pluripotent stem cells. *Nat. Methods* **7**, 989–994 (2010).
12. Villa-Diaz, L. G. et al. Synthetic polymer coatings for long-term growth of human embryonic stem cells. *Nat. Biotechnol.* **28**, 581–583 (2010).
13. Gobaa, S. et al. Artificial niche microarrays for probing single stem cell fate in high throughput. *Nat. Methods* **8**, 949–955 (2011).
14. Gefen, A. & Margulies, S. S. Are in vivo and in situ brain tissues mechanically similar? *J. Biomech.* **37**, 1339–1352 (2004).
15. Rho, J. Y., Ashman, R. B. & Turner, C. H. Young's modulus of trabecular and cortical bone material: ultrasonic and microtensile measurements. *J. Biomech.* **26**, 111–119 (1993).
16. Gilbert, P. M. et al. Substrate elasticity regulates skeletal muscle stem cell self-renewal in culture. *Science* **329**, 1078–1081 (2010).
This study demonstrated that muscle stem cells best maintained their stem cell phenotype and regenerative potential when cultured on substrates with stiffness approximating that of healthy muscle.
17. Cosgrove, B. D. et al. Rejuvenation of the muscle stem cell population restores strength to injured aged muscles. *Nat. Med.* **20**, 255–264 (2014).
18. Yang, C., Tibbitt, M. W., Basta, L. & Anseth, K. S. Mechanical memory and dosing influence stem cell fate. *Nat. Mater.* **13**, 645–652 (2014).
This study used hydrogel substrates that were dynamically softened by light to demonstrate that mesenchymal stem cells can 'remember' the stiffness of the substrates on which they were cultured.
19. Li, C. X. et al. MicroRNA-21 preserves the fibrotic mechanical memory of mesenchymal stem cells. *Nat. Mater.* **16**, 379–389 (2017).
20. Holst, J. et al. Substrate elasticity provides mechanical signals for the expansion of hemopoietic stem and progenitor cells. *Nat. Biotechnol.* **28**, 1123–1128 (2010).
21. Choi, J. S. & Harley, B. A. C. Marrow-inspired matrix cues rapidly affect early fate decisions of hematopoietic stem and progenitor cells. *Sci. Adv.* **3**, e1600455 (2017).
22. Chowdhury, F. et al. Soft substrates promote homogeneous self-renewal of embryonic stem cells via downregulating cell-matrix tractions. *PLoS ONE* **5**, e15655 (2010).
23. Madl, C. M. et al. Maintenance of neural progenitor cell stemness in 3D hydrogels requires matrix remodelling. *Nat. Mater.* **16**, 1233–1242 (2017).
These studies[23,24,42] identified mechanisms by which matrix degradation can modulate stem cell fate.
24. Gjorevski, N. et al. Designer matrices for intestinal stem cell and organoid culture. *Nature* **539**, 560–564 (2016).

25. McMurray, R. J. et al. Nanoscale surfaces for the long-term maintenance of mesenchymal stem cell phenotype and multipotency. *Nat. Mater.* **10**, 637–644 (2011).
26. Chen, W. et al. Nanotopography influences adhesion, spreading, and self-renewal of human embryonic stem cells. *ACS Nano* **6**, 4094–4103 (2012).
27. Lei, Y. & Schaffer, D. V. A fully defined and scalable 3D culture system for human pluripotent stem cell expansion and differentiation. *Proc. Natl Acad. Sci. USA* **110**, E5039–E5048 (2013).
28. Zweigerdt, R., Andree, B., Kropp, C. & Kempf, H. in *Bioreactors: Design, Operation and Novel Applications* (ed. Mandenius, C.-F.) (Wiley-VCH, Weinheim, 2016).
29. Li, Y. et al. Engineering-derived approaches for iPSC preparation, expansion, differentiation and applications. *Biofabrication* **9**, 032001 (2017).
30. Nie, Y., Bergendahl, V., Hei, D. J., Jones, J. M. & Palecek, S. P. Scalable culture and cryopreservation of human embryonic stem cells on microcarriers. *Biotechnol. Prog.* **25**, 20–31 (2009).
31. Kehoe, D. E., Jing, D., Lock, L. T. & Tzanakakis, E. S. Scalable stirred-suspension bioreactor culture of human pluripotent stem cells. *Tissue Eng. Part A* **16**, 405–421 (2010).
32. Tabata, Y., Horiguchi, I., Lutolf, M. P. & Sakai, Y. Development of bioactive hydrogel capsules for the 3D expansion of pluripotent stem cells in bioreactors. *Biomater. Sci.* **2**, 176–183 (2014).
33. Weaver, V. M. et al. Reversion of the malignant phenotype of human breast cells in three-dimensional culture and in vivo by integrin blocking antibodies. *J. Cell Biol.* **137**, 231–245 (1997).
34. Paszek, M. J. et al. Tensional homeostasis and the malignant phenotype. *Cancer Cell* **8**, 241–254 (2005).
35. Engler, A. J., Sen, S., Sweeney, H. L. & Discher, D. E. Matrix elasticity directs stem cell lineage specification. *Cell* **126**, 677–689 (2006).
This study identified substrate stiffness as a potent regulator of stem cell differentiation in 2D culture systems.
36. Huebsch, N. et al. Harnessing traction-mediated manipulation of the cell/matrix interface to control stem-cell fate. *Nat. Mater.* **9**, 518–526 (2010).
37. Saha, K. et al. Substrate modulus directs neural stem cell behavior. *Biophys. J.* **95**, 4426–4438 (2008).
38. Przybyla, L., Lakins, J. N. & Weaver, V. M. tissue mechanics orchestrate Wnt-dependent human embryonic stem cell differentiation. *Cell Stem Cell* **19**, 462–475 (2016).
39. Cameron, A. R., Frith, J. E. & Cooper-White, J. J. The influence of substrate creep on mesenchymal stem cell behaviour and phenotype. *Biomaterials* **32**, 5979–5993 (2011).
These studies[39–41] demonstrated that the viscoelastic properties of engineered extracellular matrices can modulate stem cell differentiation.
40. Chaudhuri, O. et al. Hydrogels with tunable stress relaxation regulate stem cell fate and activity. *Nat. Mater.* **15**, 326–334 (2016).
41. Das, R. K., Gocheva, V., Hammink, R., Zouani, O. F. & Rowan, A. E. Stress-stiffening-mediated stem-cell commitment switch in soft responsive hydrogels. *Nat. Mater.* **15**, 318–325 (2016).
42. Khetan, S. et al. Degradation-mediated cellular traction directs stem cell fate in covalently crosslinked three-dimensional hydrogels. *Nat. Mater.* **12**, 458–465 (2013).
43. Cosgrove, B. D. et al. N-cadherin adhesive interactions modulate matrix mechanosensing and fate commitment of mesenchymal stem cells. *Nat. Mater.* **15**, 1297–1306 (2016).
44. Kloxin, A. M., Kasko, A. M., Salinas, C. N. & Anseth, K. S. Photodegradable hydrogels for dynamic tuning of physical and chemical properties. *Science* **324**, 59–63 (2009).
45. Freeman, R. et al. Instructing cells with programmable peptide DNA hybrids. *Nat. Commun.* **8**, 15982 (2017).
46. Lam, J., Carmichael, S. T., Lowry, W. E. & Segura, T. Hydrogel design of experiments methodology to optimize hydrogel for iPSC-NPC culture. *Adv. Healthc. Mater.* **4**, 534–539 (2015).
47. Downing, T. L. et al. Biophysical regulation of epigenetic state and cell reprogramming. *Nat. Mater.* **12**, 1154–1162 (2013).
48. Caiazzo, M. et al. Defined three-dimensional microenvironments boost induction of pluripotency. *Nat. Mater.* **15**, 344–352 (2016).
49. Aguado, B. A., Mulyasasmita, W., Su, J., Lampe, K. J. & Heilshorn, S. C. Improving viability of stem cells during syringe needle flow through the design of hydrogel cell carriers. *Tissue Eng. Part A* **18**, 806–815 (2012).
This study identified shear-thinning hydrogels as material carriers to protect cells from mechanical damage during injection.
50. Cai, L., Dewi, R. E. & Heilshorn, S. C. Injectable hydrogels with in situ double network formation enhance retention of transplanted stem cells. *Adv. Funct. Mater.* **25**, 1344–1351 (2015).
51. Yan, C. et al. Injectable solid peptide hydrogel as a cell carrier: effects of shear flow on hydrogels and cell payload. *Langmuir* **28**, 6076–6087 (2012).
52. Gaffey, A. C. et al. Injectable shear-thinning hydrogels used to deliver endothelial progenitor cells, enhance cell engraftment, and improve ischemic myocardium. *J. Thorac. Cardiovasc. Surg.* **150**, 1268–1277 (2015).
53. Führmann, T. et al. Injectable hydrogel promotes early survival of induced pluripotent stem cell-derived oligodendrocytes and attenuates longterm teratoma formation in a spinal cord injury model. *Biomaterials* **83**, 23–36 (2016).
54. Vegas, A. J. et al. Long-term glycemic control using polymer-encapsulated human stem cell-derived beta cells in immune-competent mice. *Nat. Med.* **22**, 306–311 (2016).
55. Lam, J., Lowry, W. E., Carmichael, S. T. & Segura, T. Delivery of iPS-NPCs to the stroke cavity within a hyaluronic acid matrix promotes the differentiation of transplanted cells. *Adv. Funct. Mater.* **24**, 7053–7062 (2014).
56. Huebsch, N. et al. Matrix elasticity of void-forming hydrogels controls transplanted-stem-cell-mediated bone formation. *Nat. Mater.* **14**, 1269–1277 (2015).
This study demonstrated that hydrogel stiffness can modulate stem cell behaviour in vivo.
57. Darnell, M. et al. Substrate stress-relaxation regulates scaffold remodeling and bone formation in vivo. *Adv. Healthc. Mater.* **6**, 1601185 (2017).
58. Silva, G. A. et al. Selective differentiation of neural progenitor cells by high-epitope density nanofibers. *Science* **303**, 1352–1355 (2004).
59. Sleep, E. et al. Injectable biomimetic liquid crystalline scaffolds enhance muscle stem cell transplantation. *Proc. Natl Acad. Sci. USA* **114**, E7919–E7928 (2017).
60. Lovett, M., Lee, K., Edwards, A. & Kaplan, D. L. Vascularization strategies for tissue engineering. *Tissue Eng. Part B Rev.* **15**, 353–370 (2009).
61. Richardson, T. P., Peters, M. C., Ennett, A. B. & Mooney, D. J. Polymeric system for dual growth factor delivery. *Nat. Biotechnol.* **19**, 1029–1034 (2001).
62. Levenberg, S. et al. Engineering vascularized skeletal muscle tissue. *Nat. Biotechnol.* **23**, 879–884 (2005).
63. Miller, J. S. et al. Rapid casting of patterned vascular networks for perfusable engineered three-dimensional tissues. *Nat. Mater.* **11**, 768–774 (2012).
64. Arakawa, C. K., Badeau, B. A., Zheng, Y. & DeForest, C. A. Multicellular vascularized engineered tissues through user-programmable biomaterial photodegradation. *Adv. Mater.* **29**, 1703156 (2017).
65. Suuronen, E. J. et al. Functional innervation in tissue engineered models for in vitro study and testing purposes. *Toxicol. Sci.* **82**, 525–533 (2004).
66. Shvartsman, D. et al. Sustained delivery of VEGF maintains innervation and promotes reperfusion in ischemic skeletal muscles via NGF/GDNF signaling. *Mol. Ther.* **22**, 1243–1253 (2014).
67. DiMarco, R. L., Dewi, R. E., Bernal, G., Kuo, C. & Heilshorn, S. C. Protein-engineered scaffolds for in vitro 3D culture of primary adult intestinal organoids. *Biomater. Sci.* **3**, 1376–1385 (2015).
68. Cruz-Acuña, R. et al. Synthetic hydrogels for human intestinal organoid generation and colonic wound repair. *Nat. Cell Biol.* **19**, 1326–1335 (2017).
69. Lancaster, M. A. et al. Guided self-organization and cortical plate formation in human brain organoids. *Nat. Biotechnol.* **35**, 659–666 (2017).
70. Shao, Y. et al. Self-organized amniogenesis by human pluripotent stem cells in a biomimetic implantation-like niche. *Nat. Mater.* **16**, 419–425 (2017).
71. Qian, X. et al. Brain-region-specific organoids using mini-bioreactors for modeling ZIKV exposure. *Cell* **165**, 1238–1254 (2016).
72. Ma, Z. et al. Three-dimensional filamentous human diseased cardiac tissue model. *Biomaterials* **35**, 1367–1377 (2014).
73. Burridge, P. W. et al. Human induced pluripotent stem cell-derived cardiomyocytes recapitulate the predilection of breast cancer patients to doxorubicin-induced cardiotoxicity. *Nat. Med.* **22**, 547–556 (2016).
74. Lind, J. U. et al. Instrumented cardiac microphysiological devices via multimaterial three-dimensional printing. *Nat. Mater.* **16**, 303–308 (2017).
75. Ribeiro, A. J. S. et al. Contractility of single cardiomyocytes differentiated from pluripotent stem cells depends on physiological shape and substrate stiffness. *Proc. Natl Acad. Sci. USA* **112**, 12705–12710 (2015).
76. Nguyen, E. H. et al. Versatile synthetic alternatives to Matrigel for vascular toxicity screening and stem cell expansion. *Nat. Biomed. Eng.* **1**, 0096 (2017).
77. Musah, S. et al. Mature induced-pluripotent-stem-cell-derived human podocytes reconstitute kidney glomerular-capillary-wall function on a chip. *Nat. Biomed. Eng.* **1**, 0069 (2017).
78. Dekkers, J. F. et al. Characterizing responses to CFTR-modulating drugs using rectal organoids derived from subjects with cystic fibrosis. *Sci. Transl. Med.* **8**, 344ra84 (2016).
79. McLean, W. J. et al. Clonal expansion of Lgr5-positive cells from mammalian cochlea and high-purity generation of sensory hair cells. *Cell Rep.* **18**, 1917–1929 (2017).
80. Saini, A. Cystic fibrosis patients benefit from mini guts. *Cell Stem Cell* **19**, 425–427 (2016).
81. Lyon, J. Hearing restoration: a step closer? *J. Am. Med. Assoc.* **318**, 319–320 (2017).
82. Ho, A. T. V. et al. Prostaglandin E2 is essential for efficacious skeletal muscle stem-cell function, augmenting regeneration and strength. *Proc. Natl Acad. Sci. USA* **114**, 6675–6684 (2017).
83. Rosales, A. M., Vega, S. L., DelRio, F. W., Burdick, J. A. & Anseth, K. S. Hydrogels with reversible mechanics to probe dynamic cell microenvironments. *Angew. Chem. Int. Ed.* **56**, 12132–12136 (2017).
84. DeForest, C. A. & Tirrell, D. A. A photoreversible protein-patterning approach for guiding stem cell fate in three-dimensional gels. *Nat. Mater.* **14**, 523–531 (2015).
85. Lee, T. T. et al. Light-triggered in vivo activation of adhesive peptides regulates cell adhesion, inflammation and vascularization of biomaterials. *Nat. Mater.* **14**, 352–360 (2015).
This study demonstrated the feasibility of using light as a stimulus to dynamically modify biomaterial properties in vivo.
86. Cambria, E. et al. Covalent modification of synthetic hydrogels with bioactive proteins via sortase-mediated ligation. *Biomacromolecules* **16**, 2316–2326 (2015).
87. Guvendiren, M. & Burdick, J. A. Stiffening hydrogels to probe short- and long-term cellular responses to dynamic mechanics. *Nat. Commun.* **3**, 792 (2012).
88. Turner, M. et al. Toward the development of a global induced pluripotent stem cell library. *Cell Stem Cell* **13**, 382–384 (2013).
89. Rice, J. J. et al. Engineering the regenerative microenvironment with biomaterials. *Adv. Healthc. Mater.* **2**, 57–71 (2013).

90. Vishwakarma, A. et al. Engineering immunomodulatory biomaterials to tune the inflammatory response. *Trends Biotechnol.* **34**, 470–482 (2016).
91. Ali, O. A., Emerich, D., Dranoff, G. & Mooney, D. J. In situ regulation of DC subsets and T cells mediates tumor regression in mice. *Sci. Transl. Med.* **1**, 8ra19 (2009).
92. Hori, Y., Stern, P. J., Hynes, R. O. & Irvine, D. J. Engulfing tumors with synthetic extracellular matrices for cancer immunotherapy. *Biomaterials* **30**, 6757–6767 (2009).
93. Getts, D. R. et al. Microparticles bearing encephalitogenic peptides induce T-cell tolerance and ameliorate experimental autoimmune encephalomyelitis. *Nat. Biotechnol.* **30**, 1217–1224 (2012).
94. Yoon, Y. M. et al. A combination hydrogel microparticle-based vaccine prevents type 1 diabetes in non-obese diabetic mice. *Sci. Rep.* **5**, 13155 (2015).
95. Pompano, R. R. et al. Titrating T-cell epitopes within self-assembled vaccines optimizes CD4$^+$ helper T cell and antibody outputs. *Adv. Healthc. Mater.* **3**, 1898–1908 (2014).
96. Spiller, K. L. et al. Sequential delivery of immunomodulatory cytokines to facilitate the M1-to-M2 transition of macrophages and enhance vascularization of bone scaffolds. *Biomaterials* **37**, 194–207 (2015).
    This study demonstrated that regulation of the host immune response can enhance regeneration in response to engineered constructs.

Acknowledgements C.M.M. is supported by the Stanford ChEM-H Interdisciplinary Postdoctoral Training Program in Quantitative Mechanobiology. S.C.H. acknowledges support from the National Institutes of Health (NIH) (U19 AI116484 and R21 HL13804201), the National Science Foundation (DMR 1508006) and the California Institute for Regenerative Medicine (CIRM) (RT3-07948). H.M.B. acknowledges support from the NIH (R01 AG020961, R01 AR063963, R01 NS089533, and R01 HG00967401), CIRM (DISC1-10036), the American Heart Association (17CSA33590101), the Baxter Foundation, and the Li Ka Shing Foundation.

Author contributions C.M.M., S.C.H. and H.M.B. all participated in the planning, writing and editing of the article.

Competing interests The authors declare no competing interests.

Additional information
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to H.M.B.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

PERSPECTIVE

https://doi.org/10.1038/s41586-018-0068-4

Dissecting spinal cord regeneration

Michael V. Sofroniew[1]*

The inability to recover functions lost after severe spinal cord injury has been recognized for millennia and was first attributed to a failure of spinal cord neural regeneration over 100 years ago. The last forty years have seen intense research into achieving such regeneration, but in spite of conceptual advances and many reports announcing successful interventions, progress has been slow and often controversial. Here, I examine consequential advances and setbacks, and critically consider assumptions underlying certain approaches. I argue that expanding mechanistic knowledge about multiple forms of neural regeneration, why they fail and how they can restore function will resolve conceptual contentions and push the field forward.

Failure of the crushed or transected spinal cord to functionally regenerate has been recognized by military physicians for over 4,500 years[1]. Trying to understand the reason why such regeneration does not occur was among the first topics addressed by experimental neuroscientists[2] and has in recent decades been the focus of intense research. Although this long history has produced conceptual advances, the goal of practical, reproducible, functionally meaningful and clinically translatable spinal cord regeneration remains unfulfilled. Reportedly beneficial interventions have languished or failed replication. Fundamental differences of opinion have emerged regarding cellular and molecular causes for regeneration failure, sparking controversy and influencing approaches to developing interventions.

This Perspective examines advances and setbacks in spinal cord regeneration research. In a field with over 100 years of mechanistic investigation[2], I critically analyse the historical and circumstantial derivation of certain long-standing presumptions that, for better or for worse, continue to influence ideas and experimental approaches. I review evidence that regeneration comprises multiple forms of axon growth, and that diverse mechanisms contribute to failure of different forms of neural regeneration in different lesion compartments after spinal cord injury (SCI). I explore multiple ways in which regeneration can restore function, either spontaneously or after targeted interventions. I argue that neural regeneration after SCI is not a single thing, and that expanding the mechanistic knowledge about multiple forms of regeneration, why they fail and how they can restore function, will resolve conceptual contentions and push the field forward.

What is regeneration?

It is useful to begin with a consideration of terminology. Notions about what might constitute functionally meaningful spinal cord regeneration have changed markedly over time. In the nineteenth and twentieth centuries, the long-observed failure to recover after severe SCI contributed to beliefs that neural circuits were hard-wired after development, and that restoring function would require regrowth of transected axons across lesions to their original pre-injury targets. This originally narrow concept of regeneration was overturned by research revealing the enormous potential for spontaneous synapse remodelling and circuit reorganization in both healthy and injured central nervous systems (CNS). Multiple forms of axon growth and synapse remodelling can restore function after injury and fall under the broad rubric of regeneration[3–6] (Fig. 1a–g). Although some authors argue that the term regeneration should be reserved for specific forms of axon regrowth, such definitions differ among authors, fostering confusion. Because the concept of regeneration is inextricably linked to restoration of function, I opt for a broad definition that includes multiple types of function-restoring axon growth (Fig. 1). Because regeneration is not a single thing, it is critical to precisely define the types of axon growth observed or studied, ranging from structural synapse remodelling and axon sprouting above or below lesions (Fig. 1a–c, f), to long-distance axon regrowth through spared neural tissue (Fig. 1d, e) or across lesions (Fig. 1g).

How SCI lesions are organized

Different forms of regeneration are regulated by different mechanisms in different lesion compartments. SCI lesions are not homogenous. Regardless of cause, size, severity or complexity, SCI lesions exhibit three lesion compartments: the non-neural (stromal) lesion core, astrocyte scar borders and spared neural tissue that is reactive (Fig. 2a). These compartments consist of very different cell types that have entirely different roles in repair and regeneration[6]. The different cell biology and molecular mechanisms in each compartment influence axon growth or regrowth in different ways (Fig. 2b). Understanding these differences is fundamental to understanding the requirements for achieving or improving regeneration, and to designing rational, mechanistically targeted interventions.

Non-neural lesion core

Cell damage after SCI results in potentially toxic cellular debris[7]. Innate inflammation induces debris clearance, during which local stromal or mesenchymal cells, including perivascular fibroblasts, meningeal fibroblasts and pericytes, proliferate and intermingle with extravasated blood-borne cells such as fibrocytes and diverse immune cells[7]. As the cellular debris is cleared and blood-borne inflammatory cells recede, mature lesion cores (Fig. 2a) become comprised primarily of non-neural stromal cells and extracellular matrix molecules, including fibronectins, collagens, proteoglycans and laminins[6,7]. Non-neural lesion cores vary in size from small cell clusters to large areas that can be centimetres long in human SCI[8]. Other terms for the non-neural lesion core have historically included fibrotic scar, mesenchymal scar or stromal scar[2,8].

Astrocyte scar as limitans border

During debris clearance after SCI, inflammation spread is restricted by narrow astrocyte scars that form limitans borders between the

---

[1]Department of Neurobiology, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, CA, USA. *e-mail: sofroniew@mednet.ucla.edu

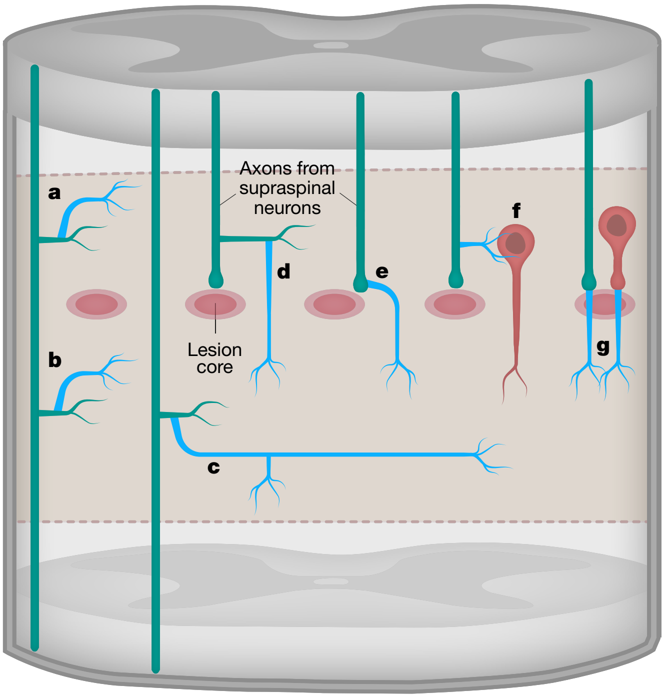


Fig. 1 | The concept of axon regeneration now encompasses multiple forms of axon growth after SCI. a, **b**, New synapse growth and remodelling by spared axons in their normal targets in response to loss of other synaptic inputs. **c**, Collateral branch sprouting of spared axons distal to incomplete SCI, and growth through spared neural tissue to reach new targets. **d**, Collateral branch sprouting of spared axons proximal to incomplete SCI, and growth through spared neural tissue to bypass scars and lesion cores and reach new targets. **e**, Regrowth from transected axon ends through spared neural tissue to bypass scars and lesion cores and reach new targets. **f**, Collateral branch sprouting of spared axons to make new contacts with propriospinal neurons that project through spared neural tissue to bypass scars and lesion cores. **g**, Axon regrowth from transected ends that passes through and across astrocyte scars and the non-neural lesion core and continues through spared neural tissue to reach new targets. Ascending axons are omitted for clarity but exhibit similar responses. Green, spared supraspinal neurons and axons; red, spared propriospinal neurons and axons; blue, new axon growth; beige, spared but reactive neural tissue of the spinal cord; dark pink, non-neural lesion cores; light pink, astrocyte scar borders.

non-neural lesion core and adjacent viable neural tissue[7,9] (Fig. 2a). These borders protect adjacent neural tissue, and their loss or dysfunction leads to spread of inflammation, larger lesions and worse functional outcome[7,9]. Astrocyte scar borders have structural and functional similarities to glia limitans borders along meningeal surfaces and blood vessels in healthy CNS that also separate neural from non-neural tissue[9]. Like other limitans borders, astrocyte scar borders are narrow and only several cells thick[9], and for this reason, the proportion of astrocyte scar to non-neural lesion core volume is small, particularly in human SCI lesions[6,8]. Astrocyte scar borders are also intermingled with reactive oligodendrocyte progenitor cells, the roles of which are not well-defined in this context[6].

Spared but reactive and reorganizing neural tissue

Astrocyte scar borders are continuous with and surrounded by extended regions of spared but reactive neural tissue that contains all neural elements required for function, including neurons, axons, dendrites, synapses, astroglia, oligodendroglia, microglia and oligodendrocyte progenitor cells[6,7] (Fig. 2a). This spared and functional neural tissue undergoes substantial circuit reorganization in the form of spontaneous synapse loss and replacement and is an important but understudied part of SCI lesions where interventions have effects and can be targeted[6].

Use of the term 'glial scar'

The term 'glial scar' derives historically from early neuroanatomists who lacked methods to distinguish among diverse glia but who nonetheless carefully discriminated two very different types of scar-forming cells in CNS lesions, those derived from non-neural mesenchymal (stromal) cells and those derived from glial cells of neural origin[2]. I argue against current usage of the term glial scar as a single conglomerate term for an 'aggregate scar' that consists of neural glia plus multiple types of non-neural cells in lesion cores[10,11]. Used in this aggregate manner, the term blurs the distinction between the many different neural and non-neural cell types that have very different roles after SCI[6,7]. I argue that the term glial scar is outmoded and ambiguous, and should be avoided in favour of referring to specific cell types in specific SCI lesion compartments.

Incomplete and complete SCI

SCI lesions vary considerably in size, complexity and functional consequences[6]. Anatomically incomplete SCI spares at least some neural connectivity across the lesion level, either directly from supraspinal sources or indirectly through propriospinal relay circuits, which may or may not lead to varying degrees of spontaneous functional recovery as circuits reorganize[6] (Fig. 3a). By contrast, anatomically complete SCI results in the complete absence of neural connectivity across the lesion level and is therefore functionally complete, and requires interventions to reconnect neural circuits across the lesion[6] (Fig. 3b). Notably, functionally complete SCI is often not anatomically complete, and may benefit from interventions that boost the reorganization and function of spared connections.

Reasons why regeneration fails

Designing experiments to improve regeneration and interpreting their results can be strongly influenced by contemporary prevailing notions about causes of regeneration failure. Notions about why regeneration fails have undergone multiple shifts over the past 100 years, and just as regeneration is not a single thing, regeneration failure has multiple causes, and these differ in different lesion compartments (Fig. 2b). A historical perspective can help us to understand the origin of current concepts and contentions.

Early mechanistic insights

Early twentieth century neuroanatomists found that axons in the adult peripheral nervous system (PNS) regenerated after injury, but axons in the CNS did not, and proposed mechanisms that are still relevant today[2]. They also showed that injured adult axons do not regrow into CNS lesions, whereas axons do regrow into live, but not dead, PNS grafts. They found that injured adult axons are attracted by extracts from live peripheral nerves soaked into porous material and implanted into PNS or CNS injury sites. Such observations contributed to Cajal's 'chemotactism' theory, that chemical attraction by 'neurotropic' substances was essential for axon growth and guidance during development and after injury, and that inadequate production of chemoattractants by injured adult CNS contributed critically to regeneration failure[2].

The inferred notion that the scar is a physical barrier

In the 1930s and 1940s, Cajal's theory of chemotactic axon guidance came under attack and was for a time dropped in favour of mechanical guidance by physical tracks and barriers[12,13]. During this period, the idea became prevalent that after CNS injury, scars physically obstructed axon regeneration, which was based on the correlative evidence that damaged axons did not regrow past neuroglial and connective-tissue scars that appeared to be obstructive[14,15]. The idea that the scar is a physical barrier was further supported by claims that pharmacologically attenuating astroglial scars enabled axon regrowth across lesions and behavioural recovery[16]. The supposed beneficial effects of these purported scar attenuators could not be replicated[17,18], but the circumstantially inferred notion of astroglial

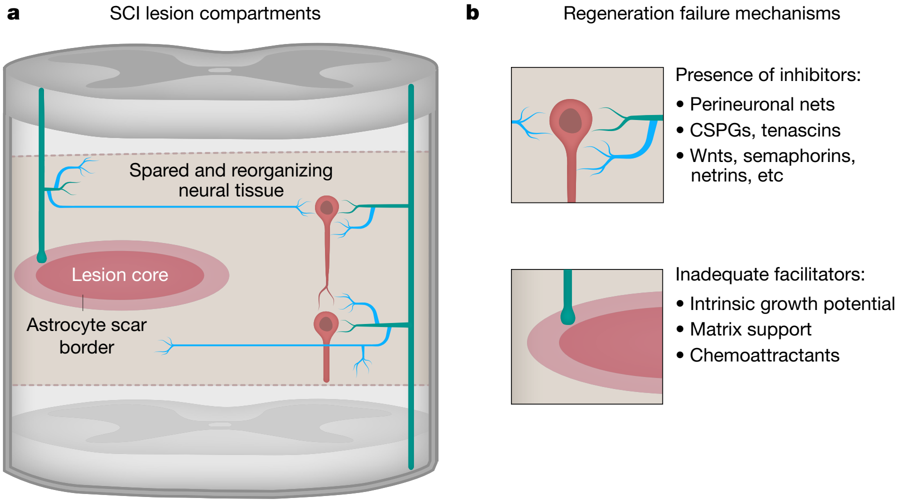


Fig. 2 | The mechanisms underlying regeneration failure vary in different SCI lesion compartments. **a**, SCI lesions exhibit three discrete compartments that consist of different cell types that have different roles in tissue repair and different forms of regeneration. The non-neural lesion core consists of stromal and residual inflammatory cells that remain after debris clearance. Astrocyte scars form limitans borders around non-neural lesion cores to restrict inflammation. Spared but reactive neural tissue contains all components of functioning neural tissue and undergoes circuit reorganization in regions that immediately surround and are continuous with the limitans borders around lesion cores. **b**, Top, in spared and reorganizing tissue, the presence of growth inhibitory mechanisms limits axon sprouting and synaptic remodelling. Bottom, in the non-neural lesion core and its limitans borders, the main cause for regeneration failure is the absence or inadequacy of multiple growth facilitators.

scar as a regeneration barrier persisted, and by the 1990s had been handed down across several generations of regeneration researchers and was firmly entrenched[14,16,19].

Absence of facilitators versus presence of inhibitors

In the 1980s, modern tract-tracing confirmed early twentieth century observations that small numbers of damaged adult CNS axons regrew into PNS grafts[20,21], leading to the exaggerated notion that all adult CNS neurons had adequate intrinsic growth capacity and that regrowth failed only because of differences in the PNS and CNS environment. Technical advances propelled a search for molecular mechanisms. Developmental neurobiology identified four types of environmental axon guidance molecules: diffusible chemoattractive, diffusible chemorepulsive, contact-attractive and contact-repulsive[22,23]. Two alternate hypotheses subsequently emerged that attributed regeneration failure either to the absence of chemoattractive facilitators or to the presence of molecular inhibitors[24]. Experimental approaches to study the potential absence of permissive chemoattractants or substrata proved technically difficult and inconclusive. By contrast, identification of molecules that were non-permissive or inhibited axon growth in vitro raised the potential that blocking inhibitors might be sufficient to achieve regeneration and recovery after injury.

CNS myelin-related growth inhibitors

Molecules associated with CNS myelin or myelin debris were among the first suggested axon growth inhibitors, including Nogo (also known as reticulon-4), myelin-associated glycoprotein, oligodendrocyte myelin glycoprotein and Nogo-receptors (NgR)[4,25]. These molecules are non-permissive to neurite outgrowth in vitro, but their roles in regeneration failure in vivo proved controversial. Early reports that blocking myelin-associated molecules with antibodies in vivo resulted in long-distance axon regeneration that mediated recovery[25–27] were called into question by findings that genetic deletion of Nogo, myelin-associated glycoprotein, oligodendrocyte myelin glycoprotein and NgR all failed to result in long-distance axon regrowth across lesions or functional recovery after SCI[28–30]. Conflicting reports and inability to reproduce original in vivo findings have been reviewed and discussed elsewhere[5,29,31]. Other studies have shown that injured adult axons from grafts[32] or after activation[33,34] regrow well along white matter tracts. Nevertheless, although NgR signalling may not have a primary role in regeneration failure across lesions, it has been compellingly implicated in regulating synapse remodelling in healthy and injured CNS[4], as discussed below.

Chondroitin sulfate proteoglycans as growth inhibitors

In the 1990s and 2000s, chondroitin sulfate proteoglycan (CSPG) production by astrocyte scars was proposed to be a principal cause for regeneration failure that acted by blocking axon regrowth across SCI lesions[35]. This notion became widespread and persists, but is also controversial. Astrocytes are not the only, or even primary producers of CSPGs in SCI lesions[36], and considerable evidence indicates that effects of modulating CSPGs after SCI are primarily related to synapse remodelling rather than to axon regrowth across scars. In uninjured CNS, neurons and astrocytes contribute CSPGs to the perineuronal net (PNN), where CSPGs and their receptors, PTPσ and NgR, regulate developmental and adult synapse plasticity[37–40]. Modulating CSPG levels or signalling in PNN can beneficially augment axon sprouting and synapse remodelling after SCI and other CNS disorders[37–40]. By contrast, multiple lines of evidence challenge the notion that the presence of CSPGs is a primary cause for failure of axons to regrow across SCI lesions. For example, CSPG digestion, genetic deletion or receptor blockade all fail to result in meaningful levels of spontaneous axon regrowth directly across astrocyte scars and non-neural lesion cores, whereas attenuating CSPGs in the PNN is sufficient to explain beneficial effects of CSPG blockade or digestion after incomplete SCI[37–45]. In keeping with this notion, aggrecan, the prototypical CSPG used to demonstrate neurite inhibition in vitro[46,47], is not present at detectable levels in astrocyte scars or the lesion core after SCI in both mice and rats, but is strongly present in the PNN of surrounding reorganizing neural tissue[36,48]. CSPG inhibition of neurite growth in vitro is relative rather than absolute, such that increasing permissive substrates, such as laminin, overrides the presence of CSPGs and vice versa[46]. This dependence on equilibrium prevents direct extrapolation to in vivo mechanisms in which both permissive and inhibitory substrates are present in SCI lesions[36], but for which the relative levels and interactions remain poorly understood. This may explain multiple in vivo studies showing that stimulated and chemoattracted axons regrow robustly through CSPGs in the astrocyte scar and lesion core in a laminin–integrin-dependent manner[36,49,50].

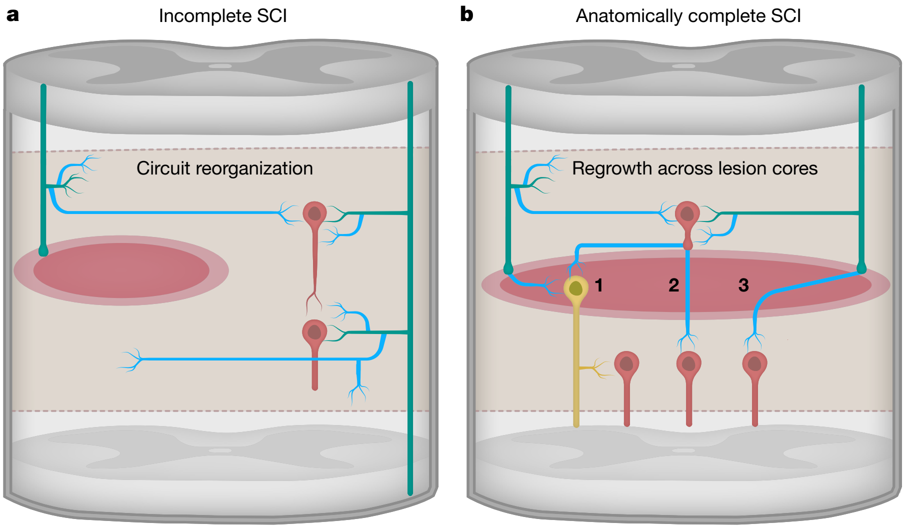


Fig. 3 | Incomplete and complete SCI lesions will benefit from targeting different forms of regeneration. **a**, Multiple forms of axon growth and synapse remodelling occur spontaneously in spared neural tissue, which after incomplete SCI can lead to circuit reorganization and formation of detour or relay pathways that restore meaningful functions. Therapies to augment such circuit reorganization can improve function. **b**, Anatomically complete SCI requires therapeutic approaches that restore connectivity of neural circuits across lesions. Approaches under investigation include: (1) Grafts of propriospinal neurons that receive host afferents and project past damaged tissue into spared neural tissue and connect into propriospinal networks or beyond. (2) Facilitating the regrowth of host propriospinal axons across damaged tissue into spared neural tissue to connect into propriospinal networks. (3) Facilitating the regrowth of host supraspinal axons across damaged tissue into spared neural tissue to connect into propriospinal networks.

Scar-forming astrocytes permit stimulated axon regrowth

For decades, barriers created by scar-forming astroglia have been proposed to be primary causes for axon regeneration failure after SCI. As discussed above, this notion was derived from correlative in vivo histology and extrapolations from in vitro findings, which suggested that scars generated by astrocytes were the primary source of putatively inhibitory CSPGs[14,16,19,35]. This view is challenged by multiple lines of evidence. Astrocytes have different roles in scar borders and spared reorganizing neural tissue[6], and understanding these distinctions can help to explain misconceptions and controversies. Hypertrophic reactive astrocytes in spared and reorganizing neural tissue produce PNN-associated CSPGs that influence synapse remodelling[38,40,51] and this role is sufficient to explain how modulating astrocyte CSPGs can influence neurite outgrowth in vitro, or augment functionally beneficial synapse remodelling in vivo. Scar-forming astrocytes form limitans borders around non-neural lesion cores that function to recruit and restrict debris-clearing inflammatory cells[7,9]. Although these limitans borders may have a cursory ‘barrier-like’ appearance, there is no rigorous causation-testing evidence that demonstrates that astrocyte borders are the primary cause of axon regrowth failure, and multiple lines of evidence from multiple laboratories argue against this notion. Neither preventing astrocyte scar formation nor removing chronic astrocyte scars leads to spontaneous regeneration of descending motor, ascending sensory or serotonin axons[36]. Appropriately stimulated CNS axons regenerate robustly in spite of astrocyte scar formation in the optic nerve[33,34] and spinal cord[36], and regrow in direct contact with scar-forming astrocytes[36,52–54]. Transgenic disruption of astrocyte scar formation does not augment, but instead significantly attenuates stimulated axon regeneration after SCI[36]. Scar-forming astrocytes express axon-growth supporting laminin[36,55] and antibody-mediated blockade of laminin–integrin binding attenuates stimulated axon regeneration after SCI[36]. Scar-forming astrocytes are newly generated cells with properties of immature astroglia[56]. Axons grow along immature astroglia during normal development[57,58], grafts of progenitor-derived immature astroglia support axon regeneration through SCI lesion cores[59–61] and spontaneous axon regeneration after SCI in lower vertebrates occurs along newly generated immature astroglia[62]. In SCI lesions, putatively inhibitory CSPGs are produced also by stromal cells, oligodendrocyte progenitor cells, inflammatory cells and other cell types, and genetically targeted astrocyte ablation does not reduce CSPGs[36]. No genetically targeted astrocyte specific loss-of-function studies report spontaneous axon regrowth across SCI lesions, and accumulating evidence indicates that such regrowth fails for other reasons.

Roles for the absence of essential growth facilitators

Lack of growth facilitators in the injured adult CNS was the first experimentally supported mechanism that was proposed for regeneration failure 100 years ago[2], and corroborating evidence has accumulated regularly. In the 1950–1990s, chemoattractive growth factors and permissive substratum were established as critical for developmental axon growth[13,63,64], and various findings indicated that similar factors are required for the regrowth of injured adult axons. Small numbers of CNS axons regrow into live PNS grafts that provide supportive substrates and chemoattractive growth factors. Regrowth does not occur into dead PNS grafts that do not produce chemoattractive factors, but can be induced by exogenously adding such factors to dead PNS grafts[20,21,65,66]. Grafts of genetically modified cells that provide a permissive substrate and chemoattractive growth factors attract injured axon regrowth across SCI lesions and beyond[67,68]. Adult CNS neurons downregulate axon growth capacity after maturity[69] and exhibit poor re-activation of intrinsic growth programs after injury[70]. CNS branches of sensory neurons are 100 times more likely to regrow into PNS grafts after SCI when manipulations are performed that reactivate the neuron-intrinsic growth capacity before SCI[71]. Genetic reactivation of neuron-intrinsic growth can reproducibly stimulate long-distance axon regrowth through adult neural tissue[33,34,72,73] but not through adult non-neural lesion core tissue[54]. Analyses of molecular mechanisms have identified multiple genetic regulators of the axon regrowth capacity of mature CNS neurons, including PTEN, mTOR, STAT3, SOCS3, MAPK, ATF3, Jun, SOX11, GAP43, arginase-1, the Klf family, MYC, RhoA, epigenetic regulators and other factors[33,34,70,74]. Combining the activation of neuron-intrinsic growth with the delivery of chemoattractive factors synergistically augments axon regrowth through SCI lesions[36]. Therefore, the absence or inadequate presence of multiple required facilitators contributes critically to regeneration failure across non-neural lesion cores and their astrocyte limitans borders (Fig. 2b).

Roles for inhibition in reorganizing grey matter

Diverse mechanisms regulate synapse formation, plasticity and remodelling, including axon-growth inhibitory signalling. During postnatal development, inhibitory PNN molecules participate in closing critical periods of synaptic plasticity, and blocking CSPG- or NgR-signalling in PNN can reopen that plasticity[4,37,38,40,75]. After SCI, blocking PNN inhibitory signalling can beneficially augment synapse remodelling and circuit reorganization[4,38,40]. Modulating other developmentally active axon-repellent cues, such as tenascins, the Wnt family, semaphorins, netrins and other factors, can also augment synapse remodelling and local axon sprouting after SCI in spared and reorganizing neural tissue in the spinal cord[37,38,40] and supraspinal centres[76] (Figs. 2b, 3a). Therefore, the effects of modulating local axon sprouting and synapse remodelling are sufficient to explain the beneficial effects of blocking such inhibitory molecules after incomplete SCI, without the need to speculatively assume that this blockade has resulted in long-distance axon regrowth directly across lesion cores in a manner that would satisfy misguided historical expectations, which unfortunately still occurs. Notably, the synapse regulatory functions of these molecules in vivo, including CSPGs, Nogo and other factors, are also sufficient to explain their non-permissive effects on neurite outgrowth in vitro.

How function can be restored

Expectations about the requirements necessary for functional restoration can influence experimental design and interpretation. Regeneration research was long swayed by historical expectations that functional recovery would require long-distance axon regrowth across lesions to reach original targets. This notion was gradually, and radically, modified by the discovery and continuously growing understanding of the substantial capacity for adult synapse remodelling and circuit reorganization in health and disease. Distinguishing between, and dissecting how, experimental interventions can induce functional recovery by augmenting the reorganization of spared circuits (Figs. 1a–f, 3a), as compared with recovery induced by achieving axon regrowth across lesions to the formation of new circuits (Figs. 1g, 3b), will be fundamental to achieving a mechanistic understanding of how to appropriately target therapeutic interventions.

Recovery via reorganization of spared circuits

The first compelling evidence for adult CNS plasticity came from injury studies, which showed that axotomy-induced loss of synaptic inputs from one source leads spontaneously to formation of new synapses by sprouting from spared adjacent terminals[77]. After SCI, synapse loss is prominent in spared but reactive neural tissue, leading to spontaneous reactive axon sprouting and synapse remodelling[78–80]. After incomplete SCI, such remodelling can restore function, as exemplified by spontaneous locomotor recovery after unilateral hemisection SCI (Brown–Sequard syndrome) in humans and experimental animals[80–82]. This recovery occurs in spite of permanent loss of descending supraspinal connections on the injured side and is mediated by spontaneous sprouting of descending supraspinal axons in corticospinal and reticulospinal tracts to cross the midline or contact propriospinal projection neurons[78–80,82,83] (Fig. 3a). SCI also triggers functionally relevant circuit reorganization in supraspinal afferent projection areas in the cerebral cortex and brainstem[73,76,84]. After SCI, synapse remodelling and circuit reorganization can be beneficially augmented by modulating inhibitory signalling[4,40,76] or increasing neuron-intrinsic growth potential[73], or by rehabilitative training combined with activation of spared circuits[85–87].

Recovery via new circuits formed across lesions

Repair of anatomically complete SCI will require formation of new circuits that connect across non-neural lesion cores and their scar borders (Figs. 1g, 3b). The fundamental question of what neural targets this new axon growth will need to reach to form new and functionally meaningful circuits remains to be answered. The capacity for spontaneous circuit reorganization discussed above suggests that recapitulating connectivity precisely as it was before injury will not be required to restore meaningful function, and that establishing short-distance axon growth across complete SCI that connects into spared propriospinal networks may instead be sufficient. The logic behind this approach derives from (i) anatomical evidence that after incomplete SCI, propriospinal networks spontaneously reorganize to form detour circuits[79,82,88] (Fig. 3a); (ii) functional evidence that such propriospinal detour circuits can mediate both locomotor function and finger dexterity[82,84]; and (iii) loss-of-function experiments, which show that detour circuits formed between supraspinal pathways and descending propriospinal neurons are required and sufficient for the recovery of voluntary locomotor function after bilateral loss of all direct projections from brain to lumbar locomotor centres[82]. These findings provide a compelling rationale for achieving short-distance axon growth from either host or grafted neurons across anatomically complete SCI lesions to connect into propriospinal circuits on the other side (Fig. 3b).

Two biological strategies to form such new connections are to (i) achieve regrowth of endogenous supraspinal, ascending or propriospinal axons across lesions that connect with spared circuits on the other side[6], or (ii) graft propriospinal neurons that receive inputs from, and send outputs to, host circuits on either side of lesions[89,90] (Fig. 3b). With regards to achieving host axon regrowth, evidence suggests that manipulating any one growth regulatory mechanism on its own will not be sufficient. For example, blocking inhibitors[5,43,44], removing putatively inhibitory astrocyte scars[36], providing permissive substrate[65,66], providing chemoattractive growth factors[36], or activating intrinsic growth programs[54], on their own all fail to achieve meaningful axon regrowth across scars and lesion cores, whereas such regrowth can be achieved by approaches that combine the activation of neuron-intrinsic growth capacity and the provision of permissive substrate and chemoattraction[6,36,70,91]. Promoting myelination to improve conduction velocity may also be required to functionalize new connections[72].

Training requirements for new circuits

Outcomes after SCI can be influenced not only by molecular or cellular interventions, but also by training and electrophysiological augmentation[85,87,92,93]. During development, acquisition of function by new circuits requires use, which generates an interplay between neural activity and synapse plasticity[94–96]. The absence of use prevents acquisition of function. Use and training are also requisite for restoring meaningful function after SCI. After unilateral hemisection SCI, locomotor function recovers spontaneously on the injured side enabled by interactions with the normally functioning contralateral limb and muscle spindle feedback[82,83]. After bilateral lesions that remove all supraspinal descending connections but spare propriospinal detour circuits, bilateral voluntary motor function does not return after simultaneous injuries where both limbs are simultaneously unable to move[82], but function can return if the injuries are separated by time so that one functional limb is always available to train the opposite side[82]. Notably, function can also return after simultaneous injuries if limb use is enabled and trained by neuroprosthetic rehabilitation and circuit activation[86], demonstrating that training-mediated use is required to functionalize new circuits formed after complete SCI.

Where we should go from here

It is unlikely that there will soon be a single intervention that is appropriate for all forms of incomplete or anatomically complete SCI. The invasive procedures that are currently required to achieve connectivity across complete lesions pose risks to functionally important residual connections that were spared in incomplete lesions, which comprise the majority of clinical SCI. Non-invasive procedures that are able to augment circuit reorganization after incomplete SCI will not restore connectivity across complete SCI. Contending with the spectrum of SCI severity will require expanding the growing mechanistic understanding that there is not a single form of spinal cord regeneration, that there is not a single reason why regeneration fails, and that different types of SCI lesions will require different ways of augmenting regeneration to improve outcome (Box 1). It is time to move past studies

BOX 1

What we have learned thus far

Regeneration is not a single thing and is controlled by multiple regulators.

Neural regeneration exhibits different forms, from synaptic remodelling to long-distance axon growth (Fig. 1). These different forms of regeneration occur in different SCI lesion compartments and are controlled by different regulators (Figs. 1, 2a), including facilitators, such as neuron-intrinsic growth programs, permissive substrates and chemoattractive molecules, and inhibitors, such as repellent substrates, chemorepellent molecules and neuron-intrinsic genetic growth suppressors.

Regeneration can fail for different reasons and can be augmented in different ways.

Because of the complex interplay of regulators, neural regeneration failure has multiple causes, including the insufficient presence of growth facilitators, or excessive presence of growth inhibitors, and their relative contributions differ in different lesion compartments (Fig. 2b). For this reason, regeneration can be augmented in multiple ways by influencing different mechanisms in different lesion compartments. Notably, spared neural tissue can contain growth facilitators, such as permissive substrates and chemoattractors, so that the inhibition of inhibitory signalling or the activation of growth programs can markedly increase synapse remodelling, axon sprouting or long-distance growth. By contrast, the non-neural lesion core lacks or is deficient in multiple growth facilitators. Thus, modulating individual growth regulators on their own, such as neutralizing inhibitors, providing chemoattractors or permissive substrates, or activating the neuron-intrinsic growth capacity, have proven to be ineffective, whereas providing combinations of permissive substrates, chemoattractive growth factors and activating intrinsic growth can result in substantial growth through non-neural lesion cores.

Function can be restored in multiple ways.

After incomplete SCI, synapse remodelling, axon sprouting and long-distance axon regrowth in spared neural tissue can all lead to reorganization of pre-existing circuits that convey functional information past lesions (Fig. 3a). After anatomically complete SCI, evidence suggests that interventions that achieve short-distance regrowth of axons across lesions may be able to enable the formation of new propriospinal circuits with the potential to restore function (Fig. 3b). Therefore, clinical interventions will need to target specific forms of regeneration in specific lesion compartments as appropriate for specific clinical contexts.

Training and use-dependent plasticity are critical for functional restoration.

After incomplete SCI, appropriate training substantially enhances functional restoration. After anatomically complete SCI, it cannot be assumed that simply achieving axon regrowth will spontaneously lead to restoration of meaningful function. Acquisition of function by new circuits requires training and use-dependent plasticity. Training benefits can be augmented by biological, electrophysiological and technological interventions.

that blur all forms of regeneration into a single undefined process, or that attempt to infer in vivo cellular and molecular mechanisms from behavioural observations, or from in vitro data, on the basis of historical expectations. It is also time to move past an excessive focus on inhibition as the primary cause for all regeneration failure, and to drop the associated expectation that blocking inhibition will on its own lead to meaningful regeneration after all forms of SCI. Modulating inhibitory signalling will have its place in specific contexts, but over-simplified and still frequently appearing statements that regeneration failure is caused simply by the inhibitory nature of SCI lesions are not accurate and can skew experimental expectations and foster misleading interpretations of data and outcomes. Instead, there is a need to expand the application to SCI research of rigorous state-of-the art technologies that dissect cellular and molecular mechanisms and test functional causality in vivo, for example by applying genetically targeted cell-type-specific loss- or gain-of-function manipulations, optogenetic technologies, trans-synaptic tract-tracing and other stringent manipulations[33,36,72,76,83,97–99]. This type of painstakingly dissected mechanistic information is essential to develop effective and translatable therapeutic strategies that reliably and reproducibly augment specific types of regeneration in functionally meaningful ways. More work is also needed on how best to use rehabilitation strategies that combine training with pharmacological and electrophysiological circuit activation to augment use-dependent plasticity that is required for biological repair to achieve maximum benefit. Together, such approaches can gradually provide therapeutic interventions tailored to augment specific types of regeneration in appropriate clinical contexts and thereby improve outcome.

Received: 1 October 2017; Accepted: 26 January 2018;
Published online 16 May 2018.

1. Wilkins, R. H. Neurosurgical Classic—XVII. *J. Neurosurg.* **21**, 240–244 (1964).
2. Ramón y Cajal, S., DeFelipe, J. & Jones, E. G. *Cajal's Degeneration and Regeneration of the Nervous System* (Oxford Univ. Press, Oxford, 1991).
3. Tuszynski, M. H. & Steward, O. Concepts and methods for the study of axonal regeneration in the CNS. *Neuron* **74**, 777–791 (2012).
4. Schwab, M. E. & Strittmatter, S. M. Nogo limits neural plasticity and recovery from injury. *Curr. Opin. Neurobiol.* **27**, 53–60 (2014).
5. Geoffroy, C. G. & Zheng, B. Myelin-associated inhibitors in axonal growth after CNS injury. *Curr. Opin. Neurobiol.* **27**, 31–38 (2014).
6. O'Shea, T. M., Burda, J. E. & Sofroniew, M. V. Cell biology of spinal cord injury and repair. *J. Clin. Invest.* **127**, 3259–3270 (2017).
7. Burda, J. E. & Sofroniew, M. V. Reactive gliosis and the multicellular response to CNS damage and disease. *Neuron* **81**, 229–248 (2014).
8. Norenberg, M. D., Smith, J. & Marcillo, A. The pathology of human spinal cord injury: defining the problems. *J. Neurotrauma* **21**, 429–440 (2004).
9. Sofroniew, M. V. Astrocyte barriers to neurotoxic inflammation. *Nat. Rev. Neurosci.* **16**, 249–263 (2015).
10. Cregg, J. M. et al. Functional regeneration beyond the glial scar. *Exp. Neurol.* **253**, 197–207 (2014).
11. Haenzi, B. & Moon, L. D. The function of FGFR1 signalling in the spinal cord: therapeutic approaches using FGFR1 ligands after spinal cord injury. *Neural Plast.* **2017**, 2740768 (2017).
12. Weiss, P. In vitro experiments on the factors determining the course of the outgrowing nerve fiber. *J. Exp. Zool.* **68**, 393–448 (1934).
13. Sperry, R. W. Chemoaffinity in the orderly growth of nerve fiber patterns and connections. *Proc. Natl Acad. Sci. USA* **50**, 703–710 (1963).
14. Sugar, O. & Gerard, R. W. Spinal cord regeneration in the rat. *J. Neurophysiol.* **3**, 1–19 (1940).
15. Brown, J. O. & McCouch, G. P. Abortive regeneration of the transected spinal cord. *J. Comp. Neurol.* **87**, 131–137 (1947).
16. Clemente, C. D. & Windle, W. F. Regeneration of severed nerve fibers in the spinal cord of the adult cat. *J. Comp. Neurol.* **101**, 691–731 (1954).
17. Arteta, J. L. Research on the regeneration of the spinal cord in the cat submitted to the action of pyrogenous substances (5 or 3895) of bacterial origin. *J. Comp. Neurol.* **105**, 171–184 (1956).
18. Matthews, M. A., St Onge, M. F., Faciane, C. L. & Gelderd, J. B. Spinal cord transection: a quantitative analysis of elements of the connective tissue matrix formed within the site of lesion following administration of piromen, cytoxan or trypsin. *Neuropathol. Appl. Neurobiol.* **5**, 161–180 (1979).
19. Reier, P. J. & Houle, J. D. The glial scar: its bearing on axonal elongation and transplantation approaches to CNS repair. *Adv. Neurol.* **47**, 87–138 (1988).
20. Richardson, P. M., McGuinness, U. M. & Aguayo, A. J. Axons from CNS neurons regenerate into PNS grafts. *Nature* **284**, 264–265 (1980).
21. David, S. & Aguayo, A. J. Axonal elongation into peripheral nervous system "bridges" after central nervous system injury in adult rats. *Science* **214**, 931–933 (1981).
22. Dodd, J. & Jessell, T. M. Axon guidance and the patterning of neuronal projections in vertebrates. *Science* **242**, 692–699 (1988).
23. Tessier-Lavigne, M. & Goodman, C. S. The molecular biology of axon guidance. *Science* **274**, 1123–1133 (1996).
24. Schwab, M. E. & Thoenen, H. Dissociated neurons regenerate into sciatic but not optic nerve explants in culture irrespective of neurotrophic factors. *J. Neurosci.* **5**, 2415–2423 (1985).
25. Schwab, M. E. Nogo and axon regeneration. *Curr. Opin. Neurobiol.* **14**, 118–124 (2004).
26. Bregman, B. S. et al. Recovery from spinal cord injury mediated by antibodies to neurite growth inhibitors. *Nature* **378**, 498–501 (1995).

27. Huang, D. W., McKerracher, L., Braun, P. E. & David, S. A therapeutic vaccine approach to stimulate axon regeneration in the adult mammalian spinal cord. *Neuron* **24**, 639–647 (1999).
28. Zheng, B. et al. Lack of enhanced spinal regeneration in Nogo-deficient mice. *Neuron* **38**, 213–224 (2003).
29. Lee, J. K. et al. Assessing spinal axon regeneration and sprouting in Nogo-, MAG-, and OMgp-deficient mice. *Neuron* **66**, 663–670 (2010).
    This study demonstrates that white matter molecules are not a primary cause for the failure of axon regrowth in vivo.
30. Zheng, B. et al. Genetic deletion of the Nogo receptor does not reduce neurite inhibition in vitro or promote corticospinal tract regeneration in vivo. *Proc. Natl Acad. Sci. USA* **102**, 1205–1210 (2005).
31. Steward, O., Zheng, B., Banos, K. & Yee, K. M. Response to: Kim et al., "Axon regeneration in young adult mice lacking Nogo-A/B." Neuron 38, 187–199. *Neuron* **54**, 191–195 (2007).
32. Davies, S. J. et al. Regeneration of adult axons in white matter tracts of the central nervous system. *Nature* **390**, 680–683 (1997).
33. Sun, F. et al. Sustained axon regeneration induced by co-deletion of PTEN and SOCS3. *Nature* **480**, 372–375 (2011).
    Ground-breaking study that shows that genetic reactivation of neuron intrinsic growth can stimulate long-distance axon growth through adult neural tissue.
34. Benowitz, L. I., He, Z. & Goldberg, J. L. Reaching the brain: advances in optic nerve regeneration. *Exp. Neurol.* **287**, 365–373 (2017).
35. Silver, J. & Miller, J. H. Regeneration beyond the glial scar. *Nat. Rev. Neurosci.* **5**, 146–156 (2004).
36. Anderson, M. A. et al. Astrocyte scar formation aids central nervous system axon regeneration. *Nature* **532**, 195–200 (2016).
    Genetic loss-of-function study that challenges the notion that scar-forming astrocytes are a primary cause for the failure of axon regrowth in vivo.
37. Wang, D. & Fawcett, J. The perineuronal net and the control of CNS plasticity. *Cell Tissue Res.* **349**, 147–160 (2012).
38. Mironova, Y. A. & Giger, R. J. Where no synapses go: gatekeepers of circuit remodeling and synaptic strength. *Trends Neurosci.* **36**, 363–373 (2013).
39. Dyck, S. M. & Karimi-Abdolrezaee, S. Chondroitin sulfate proteoglycans: key modulators in the developing and pathologic central nervous system. *Exp. Neurol.* **269**, 169–187 (2015).
40. Fawcett, J. W. The extracellular matrix in plasticity and regeneration after CNS injury and neurodegenerative disease. *Prog. Brain Res.* **218**, 213–226 (2015).
41. Bradbury, E. J. et al. Chondroitinase ABC promotes functional recovery after spinal cord injury. *Nature* **416**, 636–640 (2002).
42. Hossain-Ibrahim, M. K., Rezajooi, K., Stallcup, W. B., Lieberman, A. R. & Anderson, P. N. Analysis of axonal regeneration in the central and peripheral nervous systems of the NG2-deficient mouse. *BMC Neurosci.* **8**, 80 (2007).
43. García-Alías, G., Barkhuysen, S., Buckle, M. & Fawcett, J. W. Chondroitinase ABC treatment opens a window of opportunity for task-specific rehabilitation. *Nat. Neurosci.* **12**, 1145–1151 (2009).
    This study demonstrates that neutralizing CSPGs after SCI act by degrading PNN and augmenting synaptic remodelling without altering long-distance axon regrowth.
44. Wang, D., Ichiyama, R. M., Zhao, R., Andrews, M. R. & Fawcett, J. W. Chondroitinase combined with rehabilitation promotes recovery of forelimb function in rats with chronic spinal cord injury. *J. Neurosci.* **31**, 9332–9344 (2011).
45. Alilain, W. J., Horn, K. P., Hu, H., Dick, T. E. & Silver, J. Functional regeneration of respiratory pathways after spinal cord injury. *Nature* **475**, 196–200 (2011).
46. Tom, V. J., Steinmetz, M. P., Miller, J. H., Doller, C. M. & Silver, J. Studies on the development and behavior of the dystrophic growth cone, the hallmark of regeneration failure, in an in vitro model of the glial scar and after spinal cord injury. *J. Neurosci.* **24**, 6531–6539 (2004).
47. Lang, B. T. et al. Modulation of the proteoglycan receptor PTPσ promotes recovery after spinal cord injury. *Nature* **518**, 404–408 (2015).
48. Andrews, E. M., Richards, R. J., Yin, F. Q., Viapiano, M. S. & Jakeman, L. B. Alterations in chondroitin sulfate proteoglycan expression occur both at and far from the site of spinal contusion injury. *Exp. Neurol.* **235**, 174–187 (2012).
49. Lu, P., Jones, L. L. & Tuszynski, M. H. Axon regeneration through scars and into sites of chronic spinal cord injury. *Exp. Neurol.* **203**, 8–21 (2007).
50. Jones, L. L., Sajed, D. & Tuszynski, M. H. Axonal regeneration through regions of chondroitin sulfate proteoglycan deposition after spinal cord injury: a balance of permissiveness and inhibition. *J. Neurosci.* **23**, 9276–9288 (2003).
51. Khakh, B. S. & Sofroniew, M. V. Diversity of astrocyte functions and phenotypes in neural circuits. *Nat. Neurosci.* **18**, 942–952 (2015).
52. Kawaja, M. D. & Gage, F. H. Reactive astrocytes are substrates for the growth of adult CNS axons in the presence of elevated levels of nerve growth factor. *Neuron* **7**, 1019–1030 (1991).
53. Lee, J. K. et al. Combined genetic attenuation of myelin and semaphorin-mediated growth inhibition is insufficient to promote serotonergic axon regeneration. *J. Neurosci.* **30**, 10899–10904 (2010).
54. Zukor, K. et al. Short hairpin RNA against PTEN enhances regenerative growth of corticospinal tract axons after spinal cord injury. *J. Neurosci.* **33**, 15350–15361 (2013).
55. Frisén, J. et al. Spinal axons in central nervous system scar tissue are closely related to laminin-immunoreactive astrocytes. *Neuroscience* **65**, 293–304 (1995).
56. Wanner, I. B. et al. Glial scar borders are formed by newly proliferated, elongated astrocytes that interact to corral inflammatory and fibrotic cells via STAT3-dependent mechanisms after spinal cord injury. *J. Neurosci.* **33**, 12870–12886 (2013).
57. Mason, C. A., Edmondson, J. C. & Hatten, M. E. The extending astroglial process: development of glial cell shape, the growing tip, and interactions with neurons. *J. Neurosci.* **8**, 3124–3134 (1988).
58. Raper, J. & Mason, C. Cellular strategies of axonal pathfinding. *Cold Spring Harb. Perspect. Biol.* **2**, a001933 (2010).
59. Davies, J. E. et al. Astrocytes derived from glial-restricted precursors promote spinal cord repair. *J. Biol.* **5**, 7 (2006).
60. Shih, C. H., Lacagnina, M., Leuer-Bisciotti, K. & Pröschel, C. Astroglial-derived periostin promotes axonal regeneration after spinal cord injury. *J. Neurosci.* **34**, 2438–2443 (2014).
61. Zhang, S. et al. Thermoresponsive copolypeptide hydrogel vehicles for central nervous system cell delivery. *ACS Biomater. Sci. Eng.* **1**, 705–717 (2015).
62. Mokalled, M. H. et al. Injury-induced *ctgfa* directs glial bridging and spinal cord regeneration in zebrafish. *Science* **354**, 630–634 (2016).
63. Letourneau, P. C. Cell-to-substratum adhesion and guidance of axonal elongation. *Dev. Biol.* **44**, 92–101 (1975).
64. Campenot, R. B. Local control of neurite development by nerve growth factor. *Proc. Natl Acad. Sci. USA* **74**, 4516–4519 (1977).
65. Smith, G. V. & Stevenson, J. A. Peripheral nerve grafts lacking viable Schwann cells fail to support central nervous system axonal regeneration. *Exp. Brain Res.* **69**, 299–306 (1988).
66. Hagg, T. et al. Nerve growth factor promotes CNS cholinergic axonal regeneration into acellular peripheral nerve grafts. *Exp. Neurol.* **112**, 79–88 (1991).
67. Alto, L. T. et al. Chemotropic guidance facilitates axonal regeneration and synapse formation after spinal cord injury. *Nat. Neurosci.* **12**, 1106–1113 (2009).
    This study demonstrates that developmental chemoattractive growth factors also attract axon regrowth after SCI.
68. Deng, L. X. et al. A novel growth-promoting pathway formed by GDNF-overexpressing Schwann cells promotes propriospinal axonal regeneration, synapse formation, and partial recovery of function after spinal cord injury. *J. Neurosci.* **33**, 5655–5667 (2013).
69. Goldberg, J. L., Klassen, M. P., Hua, Y. & Barres, B. A. Amacrine-signaled loss of intrinsic axon growth ability by retinal ganglion cells. *Science* **296**, 1860–1864 (2002).
70. He, Z. & Jin, Y. Intrinsic control of axon regeneration. *Neuron* **90**, 437–451 (2016).
71. Richardson, P. M. & Issa, V. M. Peripheral injury enhances central regeneration of primary sensory neurones. *Nature* **309**, 791–793 (1984).
    This study demonstrates that reactivating intrinsic neuronal growth programs synergistically augments regrowth of injured CNS into a supportive and attractive environment.
72. Bei, F. et al. Restoration of visual function by enhancing conduction in regenerated axons. *Cell* **164**, 219–232 (2016).
73. Liu, Y. et al. A sensitized igf1 treatment restores corticospinal axon-dependent functions. *Neuron* **95**, 817–833 (2017).
74. Liu, K. et al. PTEN deletion enhances the regenerative ability of adult corticospinal neurons. *Nat. Neurosci.* **13**, 1075–1081 (2010).
75. Pizzorusso, T. et al. Reactivation of ocular dominance plasticity in the adult visual cortex. *Science* **298**, 1248–1251 (2002).
76. Hollis, E. R. II et al. Ryk controls remapping of motor cortex during functional recovery after spinal cord injury. *Nat. Neurosci.* **19**, 697–705 (2016).
77. Raisman, G. Neuronal plasticity in the septal nuclei of the adult rat. *Brain Res.* **14**, 25–48 (1969).
78. Weidner, N., Ner, A., Salimi, N. & Tuszynski, M. H. Spontaneous corticospinal axonal plasticity and functional recovery after adult central nervous system injury. *Proc. Natl Acad. Sci. USA* **98**, 3513–3518 (2001).
79. Bareyre, F. M. et al. The injured spinal cord spontaneously forms a new intraspinal circuit in adult rats. *Nat. Neurosci.* **7**, 269–277 (2004).
    This study shows that injured supraspinal axons form new connections with propriospinal neurons after SCI.
80. Rosenzweig, E. S. et al. Extensive spontaneous plasticity of corticospinal projections after primate spinal cord injury. *Nat. Neurosci.* **13**, 1505–1510 (2010).
81. Little, J. W. & Halar, E. Temporal course of motor recovery after Brown–Sequard spinal cord injuries. *Paraplegia* **23**, 39–46 (1985).
82. Courtine, G. et al. Recovery of supraspinal control of stepping via indirect propriospinal relay connections after spinal cord injury. *Nat. Med.* **14**, 69–74 (2008).
    This study shows that new relay circuits formed by propriospinal neurons are required and sufficient to relay functional information after SCI that transects all supraspinal connections to below the injury.
83. Takeoka, A., Vollenweider, I., Courtine, G. & Arber, S. Muscle spindle feedback directs locomotor recovery and circuit reorganization after spinal cord injury. *Cell* **159**, 1626–1639 (2014).
84. Nishimura, Y. et al. Time-dependent central compensatory mechanisms of finger dexterity after spinal cord injury. *Science* **318**, 1150–1155 (2007).
85. Harkema, S. et al. Effect of epidural stimulation of the lumbosacral spinal cord on voluntary movement, standing, and assisted stepping after motor complete paraplegia: a case study. *Lancet* **377**, 1938–1947 (2011).
    This study demonstrates that after functionally complete, but anatomically incomplete SCI, spared connections can persist that cannot on their own mediate voluntary function, but that can do so after electrophysiological stimulation combined with appropriate rehabilitation training.
86. van den Brand, R. et al. Restoring voluntary control of locomotion after paralyzing spinal cord injury. *Science* **336**, 1182–1185 (2012).

87. Angeli, C. A., Edgerton, V. R., Gerasimenko, Y. P. & Harkema, S. J. Altering spinal cord excitability enables voluntary movements after chronic complete paralysis in humans. *Brain* **137**, 1394–1409 (2014).
88. Alstermark, B. & Isa, T. Circuits for skilled reaching and grasping. *Annu. Rev. Neurosci.* **35**, 559–578 (2012).
89. Lu, P. et al. Long-distance growth and connectivity of neural stem cells after severe spinal cord injury. *Cell* **150**, 1264–1273 (2012).
90. Lu, P. et al. Prolonged human neural stem cell maturation supports recovery in injured rodent CNS. *J. Clin. Invest.* **127**, 3287–3299 (2017).
91. Assinck, P., Duncan, G. J., Hilton, B. J., Plemel, J. R. & Tetzlaff, W. Cell transplantation therapy for spinal cord injury. *Nat. Neurosci.* **20**, 637–647 (2017).
92. Courtine, G. et al. Transformation of nonfunctional spinal circuits into functional states after the loss of brain input. *Nat. Neurosci.* **12**, 1333–1342 (2009).
93. Capogrosso, M. et al. A brain–spine interface alleviating gait deficits after spinal cord injury in primates. *Nature* **539**, 284–288 (2016).
94. Crair, M. C., Gillespie, D. C. & Stryker, M. P. The role of visual experience in the development of columns in cat visual cortex. *Science* **279**, 566–570 (1998).
95. Inglis, F. M., Zuckerman, K. E. & Kalb, R. G. Experience-dependent development of spinal motor neurons. *Neuron* **26**, 299–305 (2000).
96. Martin, J. H., Choy, M., Pullman, S. & Meng, Z. Corticospinal system development depends on motor experience. *J. Neurosci.* **24**, 2122–2132 (2004).
97. Jin, D. et al. Restoration of skilled locomotion by sprouting corticospinal axons induced by co-deletion of PTEN and SOCS3. *Nat. Commun.* **6**, 8074 (2015).
98. Hilton, B. J. et al. Re-establishment of cortical motor output maps and spontaneous functional recovery via spared dorsolaterally projecting corticospinal neurons after dorsal column spinal cord injury in adult mice. *J. Neurosci.* **36**, 4080–4092 (2016).
99. Montgomery, K. L., Iyer, S. M., Christensen, A. J., Deisseroth, K. & Delp, S. L. Beyond the brain: optogenetic control in the spinal cord and peripheral nervous system. *Sci. Transl. Med.* **8**, 337rv5 (2016).

Acknowledgements Work in the laboratory of M.V.S. is supported by the National Institutes of Health (NS084030), the Dr Miriam and Sheldon G. Adelson Medical Foundation, the Craig H. Neilsen Foundation, Paralyzed Veterans of America, and Wings for Life.

Reviewer information Nature thanks R. Giger, B. Zheng and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Competing interests The author declares no competing interests.

Additional information
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to M.V.S.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

https://doi.org/10.1038/s41586-018-0088-0

Pancreas regeneration

Qiao Zhou[1,2]* & Douglas A. Melton[1,2,3]*

The pancreas is made from two distinct components: the exocrine pancreas, a reservoir of digestive enzymes, and the endocrine islets, the source of the vital metabolic hormone insulin. Human islets possess limited regenerative ability; loss of islet β-cells in diseases such as type 1 diabetes requires therapeutic intervention. The leading strategy for restoration of β-cell mass is through the generation and transplantation of new β-cells derived from human pluripotent stem cells. Other approaches include stimulating endogenous β-cell proliferation, reprogramming non-β-cells to β-like cells, and harvesting islets from genetically engineered animals. Together these approaches form a rich pipeline of therapeutic development for pancreatic regeneration.

The pancreas is central to the control of energy consumption and metabolism and is composed of two morphologically and functionally distinct components: the exocrine pancreas (acinar cells and ductal cells) and the endocrine pancreas (islets of Langerhans). Exocrine acinar cells produce an array of digestive enzymes, including lipases, proteinases, and amylases, which are secreted into pancreatic ducts and flow into the small intestine to break down fats, proteins, and carbohydrates for absorption. The endocrine islets represent less than 5% of total pancreatic mass but nevertheless number more than a billion cells in humans. Each of the five major types of islet cell synthesizes and secretes a principle hormone: insulin (β-cells), glucagon (α-cells), somatostatin (δ-cells), pancreatic polypeptide (PP cells), and ghrelin (ε-cells). Insulin and glucagon are released directly into the blood circulation through a dense intra-islet vascular network and have essential roles in the regulation of blood glucose levels.

Distinct diseases afflict the exocrine and endocrine pancreas. Pancreatitis and pancreatic cancers, the majority of which are ductal carcinomas, originate from the exocrine pancreas whereas diabetes and rare pancreatic neuroendocrine tumours arise from the endocrine islets. Diabetes has been estimated to afflict well over 300 million people worldwide and is a major and growing health problem in the modern world. Complications resulting from long-term diabetes include kidney failure, peripheral vascular disease, stroke, and coronary artery disease; together, these complications create enormous medical and social burdens as well as causing premature deaths. The majority of diabetic patients suffer from type 2 diabetes (T2D), a disease attributed to insulin resistance by peripheral organs including liver, fat, and muscle. Recent genetic linkage studies and histological analyses have shown that patients with T2D also have significantly fewer islet β-cells than healthy individuals[1–4]. Type 1 diabetes (T1D), which makes up about 5–10% of all diabetes cases, is an autoimmune disease in which β-cells are selectively destroyed, leading to a severe insulin deficiency that must be treated with daily insulin injections for survival. Together, these diseases account for a large and growing patient population with pancreatic β-cell deficiency.

There is a long history of investigations into pancreatic regeneration, going back nearly a century[5]. The epidemic of diabetes in recent decades has spurred numerous studies on pancreas development, homeostasis, and regeneration. Animal studies have suggested that the exocrine pancreas possesses an intrinsic capacity for regeneration and thus can make a rapid and full recovery from exocrine diseases such as acute pancreatitis. By contrast, the endocrine islets have limited regenerative capacity in adults. Indeed, it remains unclear whether the adult human pancreas can spontaneously regenerate β-cells in any physiologically meaningful way. Substantial β-cell loss therefore results in permanent endocrine deficiency and irreversible diabetes. There is an increasing consensus that a regenerative medicine approach will be helpful, even essential, in treating certain forms of diabetes including T1D and possibly the subset of T2D in which there is substantial β-cell loss. Learning how to enhance or induce the intrinsic regenerative ability of endocrine islets and devising new strategies to produce insulin-secreting β-cells will have profound implications for developing therapeutic treatment for diabetes. Here we summarize our current understanding of pancreatic endocrine and exocrine regeneration and review the different strategies for therapeutic regeneration and repair.

Regeneration of the endocrine pancreas

The majority of studies on pancreas regeneration have focused on endocrine islets, owing to their central importance in diabetes. Historically, studies of islet regeneration relied on rodent injury models, including pancreatectomy, pancreatic duct ligation, and chemical ablation of islet cells. In pancreatectomy, removal of up to 90% of the rat pancreas does not affect glucose homeostasis, suggesting a large reserve capacity, as 10% of the islet mass is sufficient to maintain blood glucose control[6–8]. By contrast, resection of 50–60% of the pancreas in humans triggers insulin-dependent diabetes[9,10]. Young rodents show tissue growth and sprouting from the cut surface after pancreatectomy[6,7]. Observations of rare samples from children also suggest tissue growth after pancreatectomy[11]. The capacity for this type of regeneration, however, declines sharply in adult animals and is absent in adult humans[8,10,12].

A second injury model used to study pancreas regeneration is duct ligation which mimics obstructive pancreatitis. Physical ligation of the pancreatic ducts causes widespread acinar cell death, but the endocrine islets are spared and no substantial endocrine regeneration is observed[13,14]. In a third injury model, pancreatic β-cells can be specifically ablated using streptozotocin (STZ) or alloxan, chemical toxins that structurally mimic glucose and are selectively imported into β-cells. Depending on drug dosage, the entire β-cell mass can be partially or nearly completely ablated in a few days. Extensive studies have found no convincing evidence for β-cell regeneration in adult animals following chemical ablation[12,15].

Despite the lack of substantial islet regeneration in injury models, islet hyperplasia is observed during pregnancy, in obesity, or under insulin resistance conditions in animal models[16–19]. For instance, mouse pancreatic β-cell mass increases by 3–5-fold during pregnancy, stimulated at least partly by the pregnancy hormones placental lactogen and prolactin, and involving signalling through serotonin, Menin, and FoxM1[20–23]. High-fat diet-induced obesity in mice is also accompanied

---

[1]Department of Stem Cell and Regenerative Biology, Harvard University, Cambridge, MA, USA. [2]Harvard Stem Cell Institute, Cambridge, MA, USA. [3]Howard Hughes Medical Institute, Chevy Chase, MD, USA. *e-mail: qiao_zhou@harvard.edu; dmelton@harvard.edu

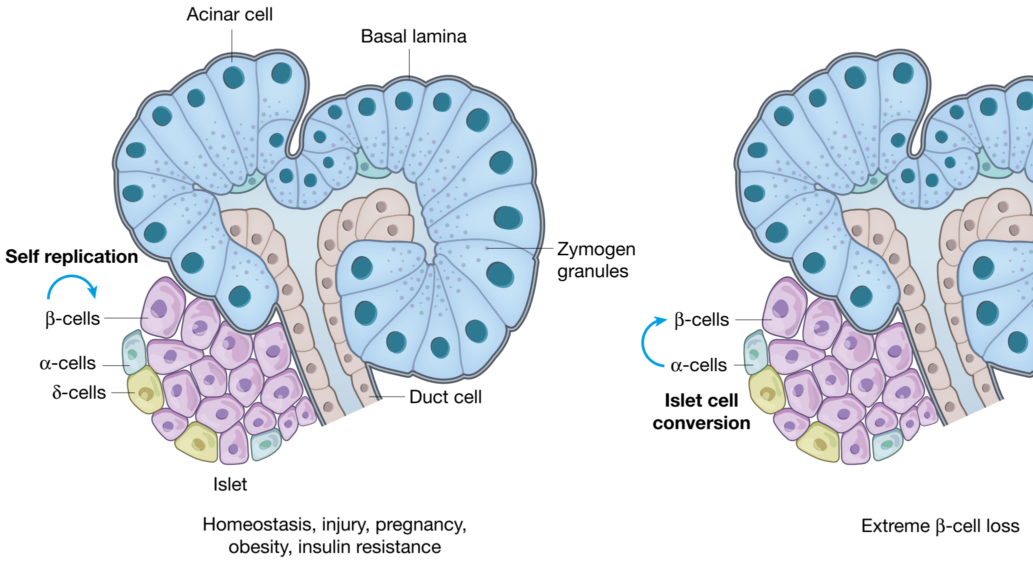


Fig. 1 | Natural regenerative responses of the endocrine pancreas. The adult endocrine pancreas (islets of Langerhans) is made up of four major endocrine cell types with each secreting a major hormone: insulin (β-cells), glucagon (α-cells), somatostatin (δ-cells), and pancreatic polypeptide (PP cells). Animal studies have shown that β-cell replication is a major mode of regeneration and repair in homeostasis, injury, pregnancy, obesity, and insulin resistance. Conversion of islet δ- and α-cells into β-cells has been reported after extreme β-cell loss using specific ablation methods in animal models. Significant regeneration of the endocrine pancreas is largely restricted to young children and young animals. Adult animals and adult humans have little, if any, ability to regenerate the endocrine pancreas.

by impressive increases in islet cell mass[24]. Experimentally induced insulin resistance, such as liver-specific knockout of insulin receptors, induces up to a tenfold increase in β-cell mass[25]. The molecular pathways that drive these increases in β-cell mass in obesity and insulin resistance have yet to be fully elucidated.

Self-replication maintains β-cell mass

The proliferative rate of β-cells is quite high in young rodents, but declines rapidly with age[26,27]. For example, one study estimated a proliferation rate of approximately 4% per day in one-month-old rats and 0.5% per day in seven-month-old rats[28]. In addition, marked islet hyperplasia can be induced in adult animals by pregnancy or obesity. What is the source of these additional islet cells? In a milestone study, genetic lineage tracing in mice using β-cell-specific drivers showed that the major mechanism for β-cell replenishment in homeostasis or after injury was replication of pre-existing β-cells[29] (Fig. 1). The role of replication is much less clear in humans, as very few replicating human β-cells (assessed by histological staining of proliferative antigens such as Ki-67 and PCNA) can be found in pancreas samples taken during autopsy of healthy, injured, pregnant, or obese adult humans[3,10,30,31].

α-cells and δ-cells may convert to β-cells

The five principle cell type of the islets, namely β-, α-, δ-, PP, and ε-cells, appear to be highly stable in normal homeostasis or in various injury models. For instance, selective ablation of β-cells with STZ or alloxan does not significantly affect the numbers or phenotypes of other islet cells. It came as a surprise that, when a diphtheria toxin-based β-cell ablation method was used to ablate more than 99% of β-cells in mice, slow but significant recovery of β-cell mass over several months was reported[32]. Lineage tracing studies suggested that the new insulin-producing cells arose from conversion of pancreatic α- or δ-cells, depending on the age of the mice[32,33]. The molecular mechanism of this conversion between islet cell types is unknown, as is whether such conversions also occur in humans. There is no clear evidence for this type of conversion in patients with T1D, but that could be either because it does not occur or because the converted cells are eliminated by the ongoing autoimmune process. Nevertheless, these studies suggest another mechanism that can potentially regenerate part of the endocrine compartment (Fig. 1).

The search for adult pancreatic stem cells

There is a long-standing hypothesis that pancreatic stem or progenitor cells might exist in the adult animal or even human pancreas[34]. This hypothesis was initially based on histological observations of single islet cells and small islets embedded in or closely associated with adult rodent and human pancreatic ducts, suggesting the emergence of new islet cells from ducts (referred to as neogenesis)[34]. However, genetic lineage-tracing studies using exocrine drivers (*Muc1*-CreER), acinar-specific drivers (*Cela*-CreER, *Ptf1a*-CreER), and duct-specific drivers (*Sox9*-CreER, *Hnf1b*-CreER)[35–39] consistently indicated rare or no contribution from the exocrine to the endocrine compartment during normal homeostasis or in various injury models. The neogenesis hypothesis has been supported by a report that, after pancreatic duct ligation in mice, a rare population of NGN3$^+$ endocrine precursor cells appeared in ductal structures[40] and observations of NGN3$^+$ cells around islets and ducts in experimental models of α-cell to β-cell transdifferentiation[41,42]. In summary, it remains unclear whether adult pancreatic stem cells exist.

Regeneration of the exocrine pancreas

The exocrine pancreas is composed of acinar cells and duct cells. The most common injury to the exocrine pancreas is pancreatitis, a painful inflammation triggered by various environmental (injury, alcohol, high fat diet and so on) or genetic (for example, cystic fibrosis) factors[43]. Understanding of exocrine damage and regeneration comes largely from rodent studies of experimental pancreatitis, the most common of which is supraphysiological stimulation of acinar secretion by caerulein, a mouse analogue of the hormone cholecystokinin[44,45]. Caerulein treatment leads to rapid apoptosis or necrosis of acinar cells and in addition, some acinar cells lose their abundant zymogen granules and shrink considerably to resemble duct cells in a process termed acinar-to-ductal metaplasia[46]. The animals recover from acute pancreatitis rapidly. Within a few weeks, the exocrine pancreas fully regains its normal cellular architecture and function. Examination of human exocrine tissues from patients with pancreatitis also shows ductal metaplasia and cell proliferation[47,48]. Although patients with acute pancreatitis can make a full recovery, it is unclear whether their exocrine pancreas undergoes similar spontaneous repair and regeneration to that seen in animal models.

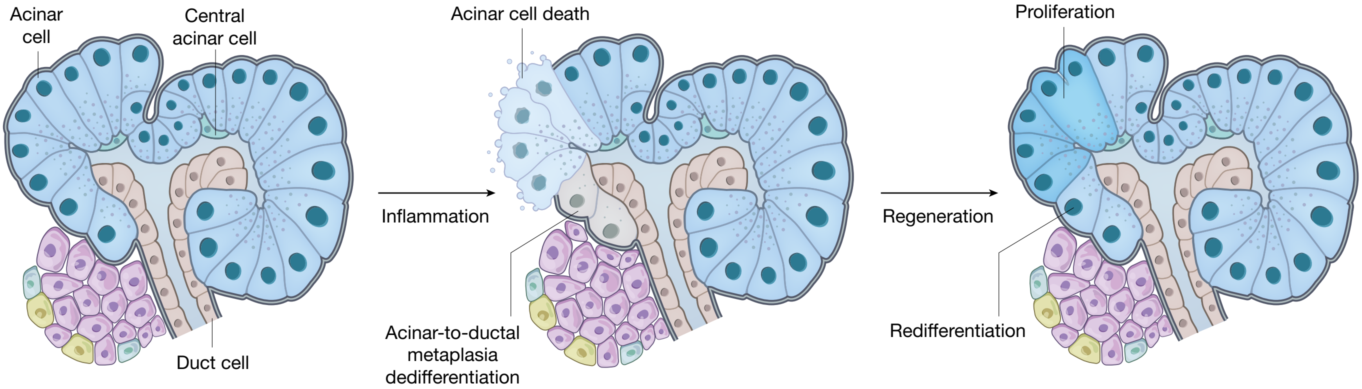


Fig. 2 | Regeneration of the exocrine pancreas. The exocrine pancreas is composed of acinar cells that synthesize and secrete digestive enzymes, ductal cells that funnel the enzymes into the small intestine, and central acinar cells. The exocrine pancreas can regenerate spontaneously and robustly in both animals and humans. Inflammatory injuries to the exocrine pancreas such as acute pancreatitis lead to acinar cell death and acinar dedifferentiation, which is characterized by degranulation and morphological transformation into duct-like cells in a process termed acinar-to-ductal metaplasia. Once inflammation subsides, acinar cells can rapidly regenerate by self-replication and possible redifferentiation of the metaplastic duct-like cells back into a normal and functional acinar state.

Two distinct modes of regeneration have been proposed to occur in models of pancreatitis[49] (Fig. 2). In the classical regeneration mode, new acinar cells are produced from proliferation of pre-existing acinar cells[35,50,51]. In the second regeneration mode, the degranulated and duct-like acinar cells are believed to 'redifferentiate' and revert back to a normal and functional acinar state. The dedifferentiated acinar cells have not been tracked with a lineage marker and their redifferentiation has been inferred by indirect means. Mechanistic studies in animal models have identified several genes and pathways required for the exocrine regenerative response in pancreatitis. Deletion of key components of the Hedgehog, Notch, and Wnt pathways from acinar cells severely disrupts exocrine regeneration[52–54], as does deletion of the acinar-restricted transcription factors NR5A2 and PTF1A[55,56]. The dedifferentiated state of acinar cells appears to represent a vulnerability in which environmental and genetic factors could conspire to induce neoplastic transformation towards the deadly pancreatic cancers[57,58]. The mechanisms that control regeneration versus neoplastic transformation are not yet understood.

Strategies to produce new endocrine islet cells

Whereas adult mouse pancreatic islets show robust regeneration under physiological challenges such as obesity, insulin resistance, or pregnancy, it is uncertain whether the adult human pancreas can deploy an adaptive regenerative response and, even if it does, these responses are evidently not able to make a significant physiological impact. At the same time, the clinical need for β-cell regeneration therapy is enormous. Approximately 2.5 million people in the USA (and more than 20 million worldwide) suffer from T1D and many millions more patients with T2D have pancreatic β-cell deficiency. Both patient populations could benefit from therapies that restore functional β-cell mass, freeing them from daily insulin injections and avoiding the serious complications that develop from imprecise dosing. The need for β-cell regeneration in patients with T1D is particularly pressing as this disease preferentially affects children and the severe lack of β-cells in T1D can cause life-threatening fluctuations in blood glucose[59,60].

Decades of clinical studies have established that cadaveric islet transplantation can be beneficial in patients with T1D, with some patients remaining free from insulin use for years[61,62]. Nevertheless, clinical cadaveric islet transplantation is used only on a small scale owing to the lack of suitable cadaveric islets and the requirement for long-term immune suppression to combat auto- and alloimmunity. To treat larger populations of patients, it would be beneficial to have a reliable and standardized source of human islets for transplantation, ideally without the need for immunosuppression. Alternatively, therapeutic interventions that stimulate endogenous islet regeneration could be used. In response to the enormous unmet medical need, several research efforts are now underway to evaluate strategies to produce new islets in vitro or stimulate islet regeneration in vivo (Fig. 3).

Differentiation of pluripotent stem cells

Decades of developmental studies in frogs, fish, and mice have mapped out the key steps and critical signalling events that lead from a fertilized egg to the formation of mature islets in early childhood[63–65]. This deep understanding of pancreatic development was put to the service of regenerative medicine in 1989, when human embryonic stem cells (hES cells) were successfully cultured and opened the door to developing methods of deriving pancreatic islets from hES cells[66]. This advance was followed in 2006 by the groundbreaking discovery that induced pluripotent stem cells (iPS cells) can be derived from somatic cells such as skin fibroblasts, providing a pathway for generating patient-specific cell products[67].

In the first major studies of deriving pancreatic endocrine cells from hES cells, a step-wise protocol was devised using combinations of signalling molecules to guide hES cell differentiation through four successive stages (definitive endoderm, pancreatic epithelium, endocrine progenitors, and β-like cells)[68,69]. The first differentiations of human stem cells into islet cells produced a population of cells with mixed hormone expression, but not mature or true human β-cells[68]. These studies, together with the decades of cellular and genetic studies of pancreatic development in animal models, created a blueprint for in vitro differentiation protocols that have been applied to pluripotent stem cells.

More recently, efforts have been directed towards the production of endocrine islets that can respond to glucose. More complicated differentiation protocols have been devised, with additional steps, optimized cocktails of inducing factors and chemicals, and the use of three-dimensional culture methods, which yield cellular clusters with remarkable morphological and functional resemblance to pancreatic islets[70,71]. Transplantation of these in vitro-derived cell clusters led to further functional maturation in vivo and robust rescue of experimental diabetes in mouse models[70,71].

In addition to endocrine islets, pancreatic progenitor cells, some of which have the capacity to produce mature hormone-producing cells, have emerged as a candidate cell therapy product[69]. Preclinical studies showed that when hES cell-derived PDX1$^+$ progenitors are transplanted into mice, some of these cells undergo growth and differentiation in vivo into functional β-cells that can reverse diabetes[69,72]. These advances have led to phase I and phase II clinical trials of pancreatic progenitors. More fully differentiated functional human islet clusters are set to enter trials in a few years. For both approaches, the in vitro-derived islet clusters contain both β-cells and other islet cell types (α-cells and δ-cells) that are known to fine-tune the function of β-cells.

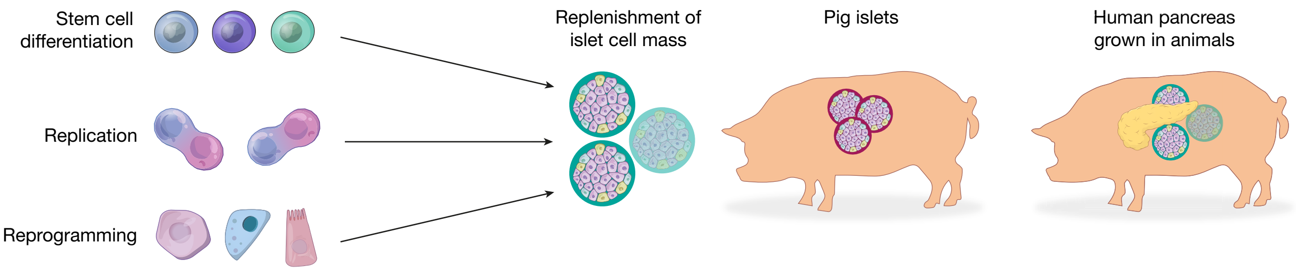


Fig. 3 | Therapeutic strategies for regeneration and repair of the endocrine pancreas. At present, the most advanced technology for making functional human insulin-secreting cells, and the only one to enter clinical trials, is derivation from human pluripotent stem cells. Other strategies include stimulating proliferation of residual β-cells in vivo, reprogramming of non-β-cells to β-like cells in vivo or in vitro, harvesting islets from genetically engineered pigs, and possibly growing entire human pancreata in animals followed by removal of human islets for transplantation.

There are additional cell types present in native islets, including vascular cells and fibroblast-like cells, and incorporating these additional cells into clusters for transplantation may offer some benefit.

Despite marked advances in producing pancreatic endocrine cells from hES cells, important challenges remain. These include perfecting the differentiation protocols for manufacturing at a large scale, eliminating unwanted cells from the final product and, of course, providing protection against immune rejection. Both allo- and autoimmune rejection can in principle be avoided by physical protection in a small device: for example, encapsulation in alginate or a more durable biomaterial. Alternatively, it may be possible to reduce the immune attack by genetic modification of the transplanted cells and/or manipulating the immune system of the recipient.

Patient-specific cell products can be derived from iPS cells, which should avoid immune rejection and be immunologically compatible with the patient from whom the iPS cell was derived. This is, of course, primarily relevant for replacing β-cells in patients with T2D, as patients with T1D suffer from autoimmunity. Manufacturing patient-specific products, however, presents its own challenges, as it will require optimization of differentiation conditions for every batch of iPS cells, adding substantial costs and operational burden to the process.

β-cell replication

Stimulating β-cell proliferation is a simple and intuitive solution to replenishing β-cell mass. Indeed, a large number of growth factors and mitogenic agents have been shown to promote β-cell proliferation in animal models. These include parathyroid hormone-related protein, hepatocyte growth factor, glucagon-like peptide, insulin-like growth factors, gastrin, epidermal growth factors, platelet-derived growth factor, adenosine kinase inhibitors, and others[16–18,73–75]. However, these agents have generally failed to promote significant proliferation of human β-cells. Substantial proliferation of human β-cells appears to occur naturally only in early childhood (mostly the first year of life)[76–79]. In all, the weight of evidence indicates that human β-cells are resistant to proliferative stimuli.

There are structural and molecular differences between mouse and human islets. For instance, β-cells are concentrated in the core of mouse islets but are more evenly distributed in human islets. Human β-cells also express several factors, such as MAFB, that are absent from mouse β-cells[80], and use GLUT1 rather than GLUT2 as the main glucose transporter[81,82]. In addition, although rodent β-cells are capable of substantial proliferation and growth in pregnancy, obesity, and insulin-resistant states, such proliferation is limited at best in adult humans. The failure of adult human β-cells to proliferate is puzzling, as they possess the necessary molecular elements that control cell cycle reentry (including cyclins, cyclin-dependent kinases (CDKs), E2F factors, and others). Direct manipulation of this molecular machinery can force human β-cells to proliferate; genetic mutations in cell cycle genes can also give rise to rare pancreatic endocrine hyperplasias such as insulinoma in humans[83–85]. Nevertheless, many of the cell cycle factors appear to be sequestered in the cytoplasm of mature β-cells[86,87]. It is unclear why this is the case or under what circumstances the cell cycle factors could be induced to traffic into the nucleus. Broad molecular and epigenetic changes occur as β-cells mature and age, with well-documented loss of EZH2 and BMI1, an increase in cell-cycle inhibitors such as $P16^{INK4a}$ and $p18^{INK4c}$, and epigenetic changes[15,88,89]. Some of these changes seem to improve β-cell function[90], but they may broadly suppress the ability of β-cells to respond to proliferative stimuli. There is longstanding evidence that insulin and glucose, both of which are elevated in obesity or insulin resistance, may directly stimulate β-cell proliferation[91–93]. But it remains unclear whether these are the key signals that drive islet hyperplasia.

A potentially important advance has come from high-throughput compound screens that identified inhibitors of dual specificity tyrosine-phosphorylation-regulated kinase 1A (DYRK1A) as reagents that can potently stimulate proliferation of cultured human β-cells in vitro and transplanted human β-cells in vivo. This provides the first concrete molecular target to manipulate human β-cell proliferation[94–96]. In addition, other pathways involved in human β-cell proliferation, such as calcineurin and SerpinB1, are being identifed[97,98]. To advance these reagents into clinics will require controlling the cell type specificity, targeting the intervention to islets, and ensuring that reagents that impinge on conserved cell cycle machineries do not raise the problem of tumour formation.

Reprogramming of non-β-cells to β-like cells

Observations of rare events in developmental cell fate changes go back many decades[99]. In the well-documented example of newt lens regeneration, removal of the lens leads to proliferation of pigmented epithelial cells surrounding the lens and regeneration of a new lens[100]. Molecular studies of master regulators of cell lineages such as MYOD further cemented the notion that powerful genetic factors could dictate cell fate choices[101]. And somatic cell nuclear transfer has demonstrated the potential of nearly every nucleus to be reprogrammed to another cell state[102]. Accordingly, there has been great interest in using master regulators of β-cell development to convert non-β-cells into insulin-producing cells. An early example is the induction of insulin expression from cultured mouse liver cells[103]. Other studies confirmed that insulin expression can be induced in non-β-cells, but these cells do not take on the morphological, molecular, and functional properties of pancreatic β-cells and are not reprogrammed to a β-like cell state[104–108].

A combinatorial screening strategy showed that a combination of three developmental regulators of β-cells, NGN3, PDX1, and MAFA (referred to as NPM factors), could efficiently convert pancreatic acinar cells into β-like cells after delivery into the adult mouse pancreas using adenoviral vectors[109]. The induced β-like cells achieved long-term stability and acquired the ability to reverse diabetes[110]. Further screening identified gastrointestinal epithelial cells as another cell type that could be converted into β-like cells[111]. Cells from the antral stomach appear to be particularly amenable to such conversion[112]. In a separate study, conditional deletion of FOXO1 from NGN3$^+$ intestinal endocrine progenitors also led to the formation of insulin-producing cells in the gut[113]. These studies together suggest that gastrointestinal epithelial cells are a potential source of functional insulin-expressing cells by reprogramming. Other examples of reprogramming mouse cells include cytokine-mediated conversion of acinar cells to insulin-expressing cells, conversion of duct cells to insulin-expressing cells by FBW7 deletion, and conversion of hepatocytes to insulin-producing cells by TGIF2[114–116].

Extreme β-cell loss can trigger spontaneous conversion of pancreatic δ- and α-cells into β-cells[32,33]. Although the molecular mechanisms of these conversion events remain unknown, genetic deletion of ARX, a regulator of α-cell development, or forced expression of PAX4, a regulator of β-cell development, can convert α-cells into β-cells in mouse models[117,118]. A unique population of insulin-producing cells at the periphery of the islets has recently been proposed to be an intermediary in the transition from α-cells to β-cells[119]. Further studies have identified γ-aminobutyric acid (GABA) signalling as a potential facilitator of the reprogramming event. Long-term GABA treatment in mice led to impressive increase in β-cell mass[42].

Despite proof-of-concept demonstrations of β-cell reprogramming in animal models, efforts to reprogram human cells have been less successful. Several studies have suggested that human α-cells in islets can be reprogrammed to become β-cells[42,120]. Changes in α-cell to β-cell ratios and the appearance of cells positive for both glucagon and insulin have provided some evidence for such conversions; however, in the absence of lineage tracing, direct evidence is still lacking. Other cell types such as pancreatic acinar cells, ductal cells, gall bladder cells, and intestinal cells have also been used to generate insulin-expressing cells, but these cells did not form long-term stable grafts after transplantation, suggesting incomplete cell fate conversion or an unstable epigenetic state[121–124]. At this point, the main challenge in translating the reprograming approach into the clinic is to define reliable methods for efficient production of human β-like cells that can develop stable and functional transplants. Besides the in vitro reprogramming approach, in vivo reprogramming in human patients targeting pancreatic α-cells, acinar cells, or gastrointestinal epithelial cells may also be feasible. It will be challenging, however, to optimize in vivo reprogramming protocols for therapeutic use in humans.

Islets from genetically engineered animals

There has been long-standing interest in islet xenotransplantation, and several exploratory clinical xenotransplantation studies with pig islets were conducted decades ago[125–127]. However, severe immune rejection of xenograft materials by the human immune system and the presence of large numbers of pig retroviruses that may jump species pose substantial obstacles. Recent advances in genetic engineering have led to reconsideration of the possibility of using organs grown in pigs. Using CRISPR–Cas9 technology, a pool of 62 known pig retroviruses was deleted from pig skin cells[128,129], which in principle could be used to make pig iPS cells, and subsequently, genetically ‘clean’ pigs as islet donors. Future clinical use of pig islets will depend on advances in encapsulation technology to protect the cells from human immune reactions while ensuring long-term survival and functionality.

Growing human pancreatic tissue in animals

The idea of growing human organs in animals for therapeutic use may seem futuristic. However, advances in stem cell technology and the identification of master regulators of organ formation have spurred efforts to explore this idea using animal models[130–132]. For example, genetic deletion of PDX1 in rats leads to specific loss of the entire pancreas due to a failure to create the embryonic pancreas. Injection of mouse embryonic stem cells into *Pdx1*$^{-/-}$ rat blastocysts created mouse–rat chimaeric animals in which all organs were made up of a mixture of mouse and rat cells except for the pancreas, which was derived from mouse cells. Thus, a mouse pancreas was grown in the body of a rat. The mouse islets from these rats can be harvested and transplanted back to diabetic mice to cure their diabetes[132]. This proof-of-concept experiment between two distinct rodent species provided a glimpse of what the future might hold for growing human organs in animal species. Nevertheless, this idea is still in its infancy. Preliminary studies have suggested that standard hES cells cannot make significant contributions to animal chimaeras[133]. Further mechanistic understanding may see the development of new methods to reduce species incompatibility and approaches that minimize or eliminate indiscriminate contribution of human cells to chimaeric animals. Aside from these technological challenges, societal consent is likely to be needed to move this technology towards clinical application.

Redifferentiating β-cells

Pancreatic β-cells become dysfunctional under a variety of stress conditions, such as prolonged hyperglycaemia and hyperlipidaemia (T2D), pancreatic inflammation due to chronic pancreatitis or pancreatic cancers (type 3c diabetes), or autoimmune-induced inflammation (T1D). Severe distress could lead to β-cell degranulation and down-regulation of β-cell genes. Recent studies have suggested that the loss of β-cell properties may represent dedifferentiation characterized by upregulation of genes that are typically expressed in embryonic islet progenitors (such as *Neurog3*)[134]. It is unclear whether dedifferentiation is a common characteristic of dysfunctional β-cells, and whether the dedifferentiation process, if it exists in humans, can be reversed. We do know that dysfunctional β-cells can recover in patients with T2D with proper management, such as diet, exercise, or intensive insulin therapy. If pharmacological means can be found to ‘redifferentiate’ the dedifferentiated β-cells, it could constitute a new therapeutic approach for diabetes and may be viewed as a distinct form of regenerative therapy, one that does not involve creation of new cells per se[135]. This therapy would be most relevant for T2D but could conceivably be helpful for early stage T1D as well.

Challenges of developing cell therapy for T1D

Developing cellular products to treat T1D faces the unique challenge of autoimmunity[59,136]. To protect the new β-cells, one can use immunosuppressants, the standard treatment for T1D patients that receive cadaveric islet transplants. However, many of these drugs are known to be toxic to β-cells, not to mention reducing the patient’s immune capacity. An alternative way to protect transplanted β-cells is encapsulation with engineered materials. Encapsulation physically separates β-cells from immune cells but it also separates β-cells from blood vessels, thus altering the kinetics of glucose sensing and oxygen and nutrient delivery, and potentially compromising the survival and function of the encapsulated cells. Innovative encapsulation materials are being developed to address these issues[137–140]. An alternative way to protect β-cells in patients with T1D is to modulate the immune system. Different immunotherapy regiments have been shown to attenuate or even completely arrest autoimmune attacks in the non-obese diabetic (NOD) mouse model[141–143]. Unfortunately, in clinical trials, these agents did not demonstrate a significant benefit for patients.

The current cell therapies for T1D require the use of encapsulation devices or immunosuppressive agents, with drawbacks and risks. Looking ahead, it would be ideal to find ways to produce islets that naturally resist autoimmunity. How might this be done? One important clue comes from studies of patients with longstanding T1D (such as the Joslin Medalist study), which made the surprising finding that a considerable number of these patients have detectable insulin production with preservation of glucose responsiveness, suggesting that some β-cells may evade autoimmunity and continue to function[144,145]. Another possibility is that new β-cells may be continuously produced in some patients. The autoimmune attack within the pancreas itself is also not uniform. Rather, some islets or even entire pancreatic lobes have been observed to escape immune destruction while surrounded by lobes depleted of β-cells[60]. These data suggest that human β-cells may be heterogeneous, and that a subpopulation of β-cells may resist the autoimmune attack. These clinical observations are consistent with accumulating evidence that mouse and human β-cells in normal islets are heterogeneous in their molecular signatures or proliferative

Table 1 | Comparison of current approaches to producing new β-cells

	Stem cell differentiation	Replication	In vivo reprogramming	Ex vivo reprogramming	Xenografts	Human pancreas
Stage of development	Phase I/II clinical	Proof-of-concept with human islets	Proof-of concept in animal models	Proof-of-concept with human cells	Proof-of-concept with pig cells	Concept development in animal models
Advantages	Unlimited supply, standardized production	No transplantation necessary	No transplantation necessary	Relatively simple production process	Unlimited supply, lower cost	Unlimited supply
Tumour risk	Moderate (teratoma)	High	Unknown, potentially high	Moderate	Low	Low
Patient specificity	Possible with iPS cells	Yes	Yes	Possible with patient cells	No	Possible with iPS cells
Issues	Complex production process	Targeted delivery required	Targeted delivery required	Stability of cell product	Strong immune reaction	Feasibility unknown, ethical concerns

potential[146–148]. A subset of β-cells has also been proposed to serve as 'hubs' for initiating pulsatile insulin release[149]. At present, it is unclear whether the β-cells found in patients with T1D are newly created in response to autoimmunity or are for some reason resistant to immune elimination. Focused studies of these β-cells in human samples and deeper understanding of the heterogeneity of human β-cells may eventually yield molecular targets that allow the production of functional insulin-secreting cells that resist autoimmunity. β-Cells derived from reprogramming of α-cells have been shown to resist autoimmunity in mouse studies, providing another potential path to study and possibly produce autoimmune-resistant β-cells[120]. Finally, it may be possible to genetically modify β-cells so that they are able to avoid detection or elimination by the autoimmune cells.

Future perspectives

We have learned a great deal about how the pancreas develops during embryogenesis and the different regenerative responses the pancreas mounts to physiological challenges and injuries. These insights are now being used to formulate regenerative strategies by differentiating stem cells, reprogramming non-β-cells, and other approaches. Fundamental research into pancreas development and homeostasis will continue to provide new insights that inspire alternative therapeutic approaches. For instance, studies of islet formation during embryogenesis may help to refine protocols for differentiating hES cells into 3D islet clusters; deeper understanding of how islets transition from the immature to mature state in postnatal development should facilitate efforts to produce functionally mature β-cells in vitro; and investigation of signals that mediate physiological expansion of β-cell mass in obesity and insulin resistance could lead to novel β-cell proliferation reagents without significant tumorigenic risks.

In spite of major advances in our understanding of pancreas regeneration, key questions remain. To name just a few: is there convincing evidence for stem cells in the adult pancreas? How heterogeneous are pancreatic β-cells in terms of function and immunological properties? What mechanisms are used by the human pancreas for natural regeneration and repair? To answer these questions, a diverse array of model systems including rodents, zebrafish, large animals, primates, and others are likely to be informative. New technologies will play an important part in advancing these studies. Single-cell analysis will provide an unprecedented view of the heterogeneity of normal and diseased islet cells, capture rare cells relevant to endocrine regeneration or immune resistance, and define transitional states from hES cells to mature β-cells or from non-β-cells to β-like cells; live cell imaging at the single-cell level will enable direct visualization of calcium waves, insulin release, and immune interactions in intact islets in vivo; humanized mouse models and human organoids could serve as surrogates to study human pancreas biology; and human genetic studies and CRISPR–Cas technology may lead to the discovery of new factors in pancreas disease and regeneration.

Clinical trials of islet cell products derived from hES cells have begun. Other approaches, including β-cell proliferation and reprogramming, may also reach the point of therapeutic development. Each approach offers certain advantages (Table 1). Beyond the safety and efficacy of these cell products, how they will fare in the T1D autoimmune environment may be a crucial determinant for their success. We are again reminded here that the ultimate goal for T1D therapy is a cell product that will naturally resist or evade autoimmunity and requires no encapsulation or immunosuppression. Close collaboration between immunologists and β-cell biologists will be necessary to make timely progress on this goal and if we succeed, it will not only benefit patients with T1D, but also offer crucial lessons in finding cures for many other autoimmune diseases.

Received: 31 October 2017; Accepted: 19 February 2018;

1. McCarthy, M. I. Genomics, type 2 diabetes, and obesity. *N. Engl. J. Med.* **363**, 2339–2350 (2010).
2. Flannick, J. & Florez, J. C. Type 2 diabetes: genetic data sharing to advance complex disease research. *Nat. Rev. Genet.* **17**, 535–549 (2016).
3. Butler, A. E. et al. β-cell deficit and increased β-cell apoptosis in humans with type 2 diabetes. *Diabetes* **52**, 102–110 (2003).
4. Rahier, J., Guiot, Y., Goebbels, R. M., Sempoux, C. & Henquin, J. C. Pancreatic β-cell mass in European subjects with type 2 diabetes. *Diabetes Obes. Metab.* **10** (Suppl. 4), 32–42 (2008).
5. Slack, J. M. Developmental biology of the pancreas. *Development* **121**, 1569–1580 (1995).
6. Lehv, M. & Fitzgerald, P. J. Pancreatic acinar cell regeneration. IV. Regeneration after resection. *Am. J. Pathol.* **53**, 513–535 (1968).
7. Bonner-Weir, S., Trent, D. F. & Weir, G. C. Partial pancreatectomy in the rat and subsequent defect in glucose-induced insulin release. *J. Clin. Invest.* **71**, 1544–1553 (1983).
8. Watanabe, H., Saito, H., Rychahou, P. G., Uchida, T. & Evers, B. M. Aging is associated with decreased pancreatic acinar cell regeneration and phosphatidylinositol 3-kinase/Akt activation. *Gastroenterology* **128**, 1391–1404 (2005).
9. Kumar, A. F., Gruessner, R. W. & Seaquist, E. R. Risk of glucose intolerance and diabetes in hemipancreatectomized donors selected for normal preoperative glucose metabolism. *Diabetes Care* **31**, 1639–1643 (2008).
10. Menge, B. A. et al. Partial pancreatectomy in adult humans does not provoke β-cell regeneration. *Diabetes* **57**, 142–149 (2008).
11. Berrocal, T., Luque, A. A., Pinilla, I. & Lassaletta, L. Pancreatic regeneration after near-total pancreatectomy in children with nesidioblastosis. *Pediatr. Radiol.* **35**, 1066–1070 (2005).
12. Rankin, M. M. & Kushner, J. A. Adaptive β-cell proliferation is severely restricted with advanced age. *Diabetes* **58**, 1365–1372 (2009).
13. Rankin, M. M. et al. β-Cells are not generated in pancreatic duct ligation-induced injury in adult mice. *Diabetes* **62**, 1634–1645 (2013).
14. Xiao, X. et al. No evidence for β cell neogenesis in murine adult pancreas. *J. Clin. Invest.* **123**, 2207–2217 (2013).
15. Tschen, S. I., Dhawan, S., Gurlo, T. & Bhushan, A. Age-dependent decline in β-cell proliferation restricts the capacity of β-cell regeneration in mice. *Diabetes* **58**, 1312–1320 (2009).
16. Mezza, T. & Kulkarni, R. N. The regulation of pre- and post-maturational plasticity of mammalian islet cell mass. *Diabetologia* **57**, 1291–1303 (2014).
17. Saunders, D. & Powers, A. C. Replicative capacity of β-cells and type 1 diabetes. *J. Autoimmun.* **71**, 59–68 (2016).
18. Wang, P. et al. Diabetes mellitus—advances and challenges in human β-cell proliferation. *Nat. Rev. Endocrinol.* **11**, 201–212 (2015).
19. Rieck, S. & Kaestner, K. H. Expansion of β-cell mass in response to pregnancy. *Trends Endocrinol. Metab.* **21**, 151–158 (2010).
20. Ernst, S., Demirci, C., Valle, S., Velazquez-Garcia, S. & Garcia-Ocaña, A. Mechanisms in the adaptation of maternal β-cells during pregnancy. *Diabetes Manag. (Lond.)* **1**, 239–248 (2011).
21. Kim, H. et al. Serotonin regulates pancreatic β-cell mass during pregnancy. *Nat. Med.* **16**, 804–808 (2010).
22. Zhang, H. et al. Gestational diabetes mellitus resulting from impaired β-cell compensation in the absence of FoxM1, a novel downstream effector of placental lactogen. *Diabetes* **59**, 143–152 (2010).
23. Karnik, S. K. et al. Menin controls growth of pancreatic β-cells in pregnant mice and promotes gestational diabetes mellitus. *Science* **318**, 806–809 (2007).

24. Kahn, S. E., Hull, R. L. & Utzschneider, K. M. Mechanisms linking obesity to insulin resistance and type 2 diabetes. *Nature* **444**, 840–846 (2006).
25. Michael, M. D. et al. Loss of insulin signaling in hepatocytes leads to severe insulin resistance and progressive hepatic dysfunction. *Mol. Cell* **6**, 87–97 (2000).
26. Finegood, D. T., Scaglia, L. & Bonner-Weir, S. Dynamics of β-cell mass in the growing rat pancreas. Estimation with a simple mathematical model. *Diabetes* **44**, 249–256 (1995).
27. Teta, M., Long, S. Y., Wartschow, L. M., Rankin, M. M. & Kushner, J. A. Very slow turnover of β-cells in aged adult mice. *Diabetes* **54**, 2557–2567 (2005).
28. Montanya, E., Nacher, V., Biarnés, M. & Soler, J. Linear correlation between beta-cell mass and body weight throughout the lifespan in Lewis rats: role of β-cell hyperplasia and hypertrophy. *Diabetes* **49**, 1341–1346 (2000).
29. Dor, Y., Brown, J., Martinez, O. I. & Melton, D. A. Adult pancreatic β-cells are formed by self-duplication rather than stem-cell differentiation. *Nature* **429**, 41–46 (2004).
This paper used genetic lineage tracing in mouse models and convincingly demonstrated β-cell replication as a major mechanism for maintaining β-cell mass in homeostasis.
30. Saisho, Y. et al. β-cell mass and turnover in humans: effects of obesity and aging. *Diabetes Care* **36**, 111–117 (2013).
31. Butler, A. E. et al. Adaptive changes in pancreatic β cell fractional area and β cell turnover in human pregnancy. *Diabetologia* **53**, 2167–2176 (2010).
32. Thorel, F. et al. Conversion of adult pancreatic α-cells to β-cells after extreme β-cell loss. *Nature* **464**, 1149–1154 (2010).
Data from this paper suggested that mouse pancreatic α-cells could naturally convert to β-cells after extreme β-cell loss.
33. Chera, S. et al. Diabetes recovery by age-dependent conversion of pancreatic δ-cells into insulin producers. *Nature* **514**, 503–507 (2014).
34. Aguayo-Mazzucato, C. & Bonner-Weir, S. Pancreatic β cell regeneration as a possible therapy for diabetes. *Cell Metab.* **27**, 57–67 (2018).
35. Desai, B. M. et al. Preexisting pancreatic acinar cells contribute to acinar cell, but not islet β cell, regeneration. *J. Clin. Invest.* **117**, 971–977 (2007).
36. Kopp, J. L. et al. Sox9$^+$ ductal cells are multipotent progenitors throughout development but do not produce new endocrine cells in the normal or injured adult pancreas. *Development* **138**, 653–665 (2011).
37. Solar, M. et al. Pancreatic exocrine duct cells give rise to insulin-producing β cells during embryogenesis but not after birth. *Dev. Cell* **17**, 849–860 (2009).
38. Pan, F. C. et al. Spatiotemporal patterns of multipotentiality in Ptf1a-expressing cells during pancreas organogenesis and injury-induced facultative restoration. *Development* **140**, 751–764 (2013).
39. Kopinke, D. & Murtaugh, L. C. Exocrine-to-endocrine differentiation is detectable only prior to birth in the uninjured mouse pancreas. *BMC Dev. Biol.* **10**, 38 (2010).
40. Xu, X. et al. β cells can be generated from endogenous progenitors in injured adult mouse pancreas. *Cell* **132**, 197–207 (2008).
41. Al-Hasani, K. et al. Adult duct-lining cells can reprogram into β-like cells able to counter repeated cycles of toxin-induced diabetes. *Dev. Cell* **26**, 86–100 (2013).
42. Ben-Othman, N. et al. Long-term GABA administration induces α-cell-mediated β-like cell neogenesis. *Cell* **168**, 73–85 (2017).
43. Lowenfels, A. B., Sullivan, T., Fiorianti, J. & Maisonneuve, P. The epidemiology and impact of pancreatic diseases in the United States. *Curr. Gastroenterol. Rep.* **7**, 90–95 (2005).
44. Willemer, S., Elsässer, H. P. & Adler, G. Hormone-induced pancreatitis. *Eur. Surg. Res.* **24** (Suppl. 1), 29–39 (1992).
45. Lerch, M. M. & Gorelick, F. S. Models of acute and chronic pancreatitis. *Gastroenterology* **144**, 1180–1193 (2013).
46. Bockman, D. E. Morphology of the exocrine pancreas related to pancreatitis. *Microsc. Res. Tech.* **37**, 509–519 (1997).
47. Bockman, D. E., Boydston, W. R. & Anderson, M. C. Origin of tubular complexes in human chronic pancreatitis. *Am. J. Surg.* **144**, 243–249 (1982).
48. Willemer, S. & Adler, G. Histochemical and ultrastructural characteristics of tubular complexes in human acute pancreatitis. *Dig. Dis. Sci.* **34**, 46–55 (1989).
49. Murtaugh, L. C. & Keefe, M. D. Regeneration and repair of the exocrine pancreas. *Annu. Rev. Physiol.* **77**, 229–249 (2015).
50. Blaine, S. A. et al. Adult pancreatic acinar cells give rise to ducts but not endocrine cells in response to growth factor signaling. *Development* **137**, 2289–2296 (2010).
51. Strobel, O. et al. In vivo lineage tracing defines the role of acinar-to-ductal transdifferentiation in inflammatory ductal metaplasia. *Gastroenterology* **133**, 1999–2009 (2007).
52. Morris, J. P. IV, Cano, D. A., Sekine, S., Wang, S. C. & Hebrok, M. β-catenin blocks Kras-dependent reprogramming of acini into pancreatic cancer precursor lesions in mice. *J. Clin. Invest.* **120**, 508–520 (2010).
53. Fendrich, V. et al. Hedgehog signaling is required for effective regeneration of exocrine pancreas. *Gastroenterology* **135**, 621–631 (2008).
54. Siveke, J. T. et al. Notch signaling is required for exocrine regeneration after acute pancreatitis. *Gastroenterology* **134**, 544–555 (2008).
55. Hoang, C. Q. et al. Transcriptional maintenance of pancreatic acinar identity, differentiation, and homeostasis by PTF1A. *Mol. Cell. Biol.* **36**, 3033–3047 (2016).
56. von Figura, G., Morris, J. P. IV, Wright, C. V. & Hebrok, M. Nr5a2 maintains acinar cell differentiation and constrains oncogenic Kras-mediated pancreatic neoplastic initiation. *Gut* **63**, 656–664 (2014).
57. Kopp, J. L. et al. Identification of Sox9-dependent acinar-to-ductal reprogramming as the principal mechanism for initiation of pancreatic ductal adenocarcinoma. *Cancer Cell* **22**, 737–750 (2012).
58. Stanger, B. Z. & Hebrok, M. Control of cell identity in pancreas development and regeneration. *Gastroenterology* **144**, 1170–1179 (2013).
59. Bluestone, J. A., Herold, K. & Eisenbarth, G. Genetics, pathogenesis and clinical interventions in type 1 diabetes. *Nature* **464**, 1293–1300 (2010).
60. Atkinson, M. A. et al. How does type 1 diabetes develop? The notion of homicide or β-cell suicide revisited. *Diabetes* **60**, 1370–1379 (2011).
61. Lakey, J. R., Mirbolooki, M. & Shapiro, A. M. Current status of clinical islet cell transplantation. *Methods Mol. Biol.* **333**, 47–104 (2006).
62. Hering, B. J. et al. Phase 3 trial of transplantation of human islets in type 1 diabetes complicated by severe hypoglycemia. *Diabetes Care* **39**, 1230–1240 (2016).
63. Arda, H. E., Benitez, C. M. & Kim, S. K. Gene regulatory networks governing pancreas development. *Dev. Cell* **25**, 5–13 (2013).
64. McCracken, K. W. & Wells, J. M. Molecular pathways controlling pancreas induction. *Semin. Cell Dev. Biol.* **23**, 656–662 (2012).
65. Murtaugh, L. C. & Melton, D. A. Genes, signals, and lineages in pancreas development. *Annu. Rev. Cell Dev. Biol.* **19**, 71–89 (2003).
66. Thomson, J. A. et al. Embryonic stem cell lines derived from human blastocysts. *Science* **282**, 1145–1147 (1998).
67. Takahashi, K. et al. Induction of pluripotent stem cells from adult human fibroblasts by defined factors. *Cell* **131**, 861–872 (2007).
68. D'Amour, K. A. et al. Production of pancreatic hormone-expressing endocrine cells from human embryonic stem cells. *Nat. Biotechnol*. 24, 1392–1401 (2006).
Refs 68 and 69 were among the first to report differentiation of hES cells toward pancreatic endocrine progenitors and islet cells.
69. Kroon, E. et al. Pancreatic endoderm derived from human embryonic stem cells generates glucose-responsive insulin-secreting cells in vivo. *Nat. Biotechnol.* **26**, 443–452 (2008).
70. Pagliuca, F. W. et al. Generation of functional human pancreatic β cells in vitro. *Cell* **159**, 428–439 (2014).
Refs 70 and 71 reported successful generation of glucose-sensitive islet clusters by in vitro differentiation of hES and iPS cells.
71. Rezania, A. et al. Reversal of diabetes with insulin-producing cells derived in vitro from human pluripotent stem cells. *Nat. Biotechnol.* **32**, 1121–1133 (2014).
72. Szot, G. L. et al. Tolerance induction and reversal of diabetes in mice transplanted with human embryonic stem cell-derived pancreatic endoderm. *Cell Stem Cell* **16**, 148–157 (2015).
73. Andersson, O. et al. Adenosine signaling promotes regeneration of pancreatic β cells in vivo. *Cell Metab.* **15**, 885–894 (2012).
74. Schulz, N. et al. Critical role for adenosine receptor A2a in β-cell proliferation. *Mol. Metab.* **5**, 1138–1146 (2016).
75. Annes, J. P. et al. Adenosine kinase inhibition selectively promotes rodent and porcine islet β-cell replication. *Proc. Natl Acad. Sci. USA* **109**, 3915–3920 (2012).
76. Kassem, S. A., Ariel, I., Thornton, P. S., Scheimberg, I. & Glaser, B. Beta-cell proliferation and apoptosis in the developing normal human pancreas and in hyperinsulinism of infancy. *Diabetes* **49**, 1325–1333 (2000).
77. Meier, J. J. et al. β-cell replication is the primary mechanism subserving the postnatal expansion of β-cell mass in humans. *Diabetes* **57**, 1584–1594 (2008).
78. Köhler, C. U. et al. Cell cycle control of β-cell replication in the prenatal and postnatal human pancreas. *Am. J. Physiol. Endocrinol. Metab.* **300**, E221–E230 (2011).
79. Gregg, B. E. et al. Formation of a human β-cell population within pancreatic islets is set early in life. *J. Clin. Endocrinol. Metab.* **97**, 3197–3206 (2012).
80. Dai, C. et al. Islet-enriched gene expression and glucose-induced insulin secretion in human and mouse islets. *Diabetologia* **55**, 707–718 (2012).
81. De Vos, A. et al. Human and rat β cells differ in glucose transporter but not in glucokinase gene expression. *J. Clin. Invest.* **96**, 2489–2495 (1995).
82. Ferrer, J., Benito, C. & Gomis, R. Pancreatic islet GLUT2 glucose transporter mRNA and protein expression in humans with and without NIDDM. *Diabetes* **44**, 1369–1374 (1995).
83. Kulkarni, R. N., Mizrachi, E. B., Ocana, A. G. & Stewart, A. F. Human β-cell proliferation and intracellular signaling: driving in the dark without a road map. *Diabetes* **61**, 2205–2213 (2012).
84. Bernal-Mizrachi, E. et al. Human β-cell proliferation and intracellular signaling part 2: still driving in the dark without a road map. *Diabetes* **63**, 819–831 (2014).
85. Stewart, A. F. et al. Human β-cell proliferation and intracellular signaling: part 3. *Diabetes* **64**, 1872–1885 (2015).
86. Fiaschi-Taesch, N. M. et al. Human pancreatic β-cell G1/S molecule cell cycle atlas. *Diabetes* **62**, 2450–2459 (2013).
87. Fiaschi-Taesch, N. M. et al. Cytoplasmic-nuclear trafficking of G1/S cell cycle molecules and adult human β-cell replication: a revised model of human β-cell G1/S control. *Diabetes* **62**, 2460–2470 (2013).
88. Krishnamurthy, J. et al. p16^{INK4a} induces an age-dependent decline in islet regenerative potential. *Nature* **443**, 453–457 (2006).
89. Chen, H. et al. Polycomb protein Ezh2 regulates pancreatic β-cell Ink4a/Arf expression and regeneration in diabetes mellitus. *Genes Dev.* **23**, 975–985 (2009).
90. Helman, A. et al. p16^{Ink4a}-induced senescence of pancreatic beta cells enhances insulin secretion. *Nat. Med.* **22**, 412–420 (2016).
91. Kulkarni, R. N. New insights into the roles of insulin/IGF-I in the development and maintenance of β-cell mass. *Rev. Endocr. Metab. Disord.* **6**, 199–210 (2005).
92. Dadon, D. et al. Glucose metabolism: key endogenous regulator of β-cell replication and survival. *Diabetes Obes. Metab.* **14** (Suppl 3), 101–108 (2012).

93. Stamateris, R. E. et al. Glucose induces mouse β-cell proliferation via IRS2, MTOR, and cyclin D2 but not the insulin receptor. *Diabetes* **65**, 981–995 (2016).
94. Wang, P. et al. A high-throughput chemical screen reveals that harmine-mediated inhibition of DYRK1A increases human pancreatic β cell replication. *Nat. Med.* **21**, 383–388 (2015).
    Refs 94, 95 and 96 identified DRYK1 inhibitors as reagents that stimulate human β-cell proliferation.
95. Dirice, E. et al. Inhibition of DYRK1A stimulates human β-cell proliferation. *Diabetes* **65**, 1660–1671 (2016).
96. Shen, W. et al. Inhibition of DYRK1A and GSK3B induces human β-cell proliferation. *Nat. Commun.* **6**, 8372 (2015).
97. El Ouaamari, A. et al. SerpinB1 promotes pancreatic β cell proliferation. *Cell Metab.* **23**, 194–205 (2016).
98. Dai, C. et al. Age-dependent human β cell proliferation induced by glucagon-like peptide 1 and calcineurin signaling. *J. Clin. Invest.* **127**, 3835–3844 (2017).
99. Slack, J. M. Metaplasia and transdifferentiation: from pure biology to the clinic. *Nat. Rev. Mol. Cell Biol.* **8**, 369–378 (2007).
100. Eguchi, G. & Okada, T. S. Differentiation of lens tissue from the progeny of chick retinal pigment cells cultured in vitro: a demonstration of a switch of cell types in clonal cell culture. *Proc. Natl Acad. Sci. USA* **70**, 1495–1499 (1973).
101. Choi, J. et al. MyoD converts primary dermal fibroblasts, chondroblasts, smooth muscle, and retinal pigmented epithelial cells into striated mononucleated myoblasts and multinucleated myotubes. *Proc. Natl Acad. Sci. USA* **87**, 7988–7992 (1990).
102. Gurdon, J. B. From nuclear transfer to nuclear reprogramming: the reversal of cell differentiation. *Annu. Rev. Cell Dev. Biol.* **22**, 1–22 (2006).
103. Ferber, S. et al. Pancreatic and duodenal homeobox gene 1 induces expression of insulin genes in liver and ameliorates streptozotocin-induced hyperglycemia. *Nat. Med.* **6**, 568–572 (2000).
104. Heremans, Y. et al. Recapitulation of embryonic neuroendocrine differentiation in adult human pancreatic duct cells expressing neurogenin 3. *J. Cell Biol.* **159**, 303–312 (2002).
105. Gasa, R. et al. Proendocrine genes coordinate the pancreatic islet differentiation program in vitro. *Proc. Natl Acad. Sci. USA* **101**, 13245–13250 (2004).
106. Kaneto, H. et al. PDX-1/VP16 fusion protein, together with NeuroD or Ngn3, markedly induces insulin gene transcription and ameliorates glucose tolerance. *Diabetes* **54**, 1009–1022 (2005).
107. Minami, K., Okano, H., Okumachi, A. & Seino, S. Role of cadherin-mediated cell–cell adhesion in pancreatic exocrine-to-endocrine transdifferentiation. *J. Biol. Chem.* **283**, 13753–13761 (2008).
108. Baeyens, L. et al. In vitro generation of insulin-producing β cells from adult exocrine pancreatic cells. *Diabetologia* **48**, 49–57 (2005).
109. Zhou, Q., Brown, J., Kanarek, A., Rajagopal, J. & Melton, D. A. In vivo reprogramming of adult pancreatic exocrine cells to β-cells. *Nature* **455**, 627–632 (2008).
     This paper showed that it is possible to directly convert pancreatic acinar cells to β-like cells in adult mice.
110. Li, W. et al. Long-term persistence and development of induced pancreatic β cells generated by lineage conversion of acinar cells. *Nat. Biotechnol.* **32**, 1223–1230 (2014).
111. Chen, Y. J. et al. De novo formation of insulin-producing "neo-β cell islets" from intestinal crypts. *Cell Reports* **6**, 1046–1058 (2014).
112. Ariyachet, C. et al. Reprogrammed stomach tissue as a renewable source of functional β cells for blood glucose regulation. *Cell Stem Cell* **18**, 410–421 (2016).
113. Talchai, C., Xuan, S., Kitamura, T., DePinho, R. A. & Accili, D. Generation of functional insulin-producing cells in the gut by Foxo1 ablation. *Nat Genet.* **44**, 406–412 (2012).
114. Baeyens, L. et al. Transient cytokine treatment induces acinar cell reprogramming and regenerates functional beta cell mass in diabetic mice. *Nat. Biotechnol.* **32**, 76–83 (2014).
115. Sancho, R., Gruber, R., Gu, G. & Behrens, A. Loss of Fbw7 reprograms adult pancreatic ductal cells into α, δ, and β cells. *Cell Stem Cell* **15**, 139–153 (2014).
116. Cerdá-Esteban, N. et al. Stepwise reprogramming of liver cells to a pancreas progenitor state by the transcriptional regulator Tgif2. *Nat. Commun.* **8**, 14127 (2017).
117. Courtney, M. et al. The inactivation of Arx in pancreatic α-cells triggers their neogenesis and conversion into functional β-like cells. *PLoS Genet.* **9**, e1003934 (2013).
118. Collombat, P. et al. The ectopic expression of Pax4 in the mouse pancreas converts progenitor cells into α and subsequently β cells. *Cell* **138**, 449–462 (2009).
119. van der Meulen, T. et al. Virgin β cells persist throughout life at a neogenic niche within pancreatic islets. *Cell Metab.* **25**, 911–926 (2017).
120. Xiao, X. et al. Endogenous reprogramming of α cells into β cells, induced by viral gene therapy, reverses autoimmune diabetes. *Cell Stem Cell* **22**, 78–90 (2018).
121. Lee, J. et al. Expansion and conversion of human pancreatic ductal cells into insulin-secreting endocrine cells. *eLife* **2**, e00940 (2013).
122. Bouchi, R. et al. FOXO1 inhibition yields functional insulin-producing cells in human gut organoid cultures. *Nat. Commun.* **5**, 4242 (2014).
123. Galivo, F. et al. Reprogramming human gallbladder cells into insulin-producing β-like cells. *PLoS ONE* **12**, e0181812 (2017).
124. Lemper, M. et al. Reprogramming of human pancreatic exocrine cells to β-like cells. *Cell Death Differ.* **22**, 1117–1130 (2015).
125. Sun, Y., Ma, X., Zhou, D., Vacek, I. & Sun, A. M. Normalization of diabetes in spontaneously diabetic cynomologus monkeys by xenografts of microencapsulated porcine islets without immunosuppression. *J. Clin. Invest.* **98**, 1417–1422 (1996).
126. Dufrane, D., Goebbels, R. M., Saliez, A., Guiot, Y. & Gianello, P. Six-month survival of microencapsulated pig islets and alginate biocompatibility in primates: proof of concept. *Transplantation* **81**, 1345–1353 (2006).
127. Elliott, R. B. Towards xenotransplantation of pig islets in the clinic. *Curr. Opin. Organ Transplant.* **16**, 195–200 (2011).
128. Niu, D. et al. Inactivation of porcine endogenous retrovirus in pigs using CRISPR–Cas9. *Science* **357**, 1303–1307 (2017).
129. Yang, L. et al. Genome-wide inactivation of porcine endogenous retroviruses (PERVs). *Science* **350**, 1101–1104 (2015).
130. Kobayashi, T. et al. Generation of rat pancreas in mouse by interspecific blastocyst injection of pluripotent stem cells. *Cell* **142**, 787–799 (2010).
131. Rashid, T., Kobayashi, T. & Nakauchi, H. Revisiting the flight of Icarus: making human organs from PSCs with large animal chimeras. *Cell Stem Cell* **15**, 406–409 (2014).
132. Yamaguchi, T. et al. Interspecies organogenesis generates autologous functional islets. *Nature* **542**, 191–196 (2017).
     This paper demonstrated the feasibility of harvesting interspecies-derived islets to control diabetes with rodent models.
133. Wu, J. et al. Interspecies chimerism with mammalian pluripotent stem cells. *Cell* **168**, 473–486 (2017).
134. Talchai, C., Xuan, S., Lin, H. V., Sussel, L. & Accili, D. Pancreatic β cell dedifferentiation as a mechanism of diabetic β cell failure. *Cell* **150**, 1223–1234 (2012).
     This paper suggested that dedifferentiation is a potential major mechanism for β-cell failure in T2D.
135. Accili, D. et al. When β-cells fail: lessons from dedifferentiation. *Diabetes Obes. Metab.* **18** (Suppl. 1), 117–122 (2016).
136. Orlando, G. et al. Cell replacement strategies aimed at reconstitution of the β-cell compartment in type 1 diabetes. *Diabetes* **63**, 1433–1444 (2014).
137. Vegas, A. J. et al. Long-term glycemic control using polymer-encapsulated human stem cell-derived β cells in immune-competent mice. *Nat. Med.* **22**, 306–311 (2016).
138. An, D. et al. Designing a retrievable and scalable cell encapsulation device for potential treatment of type 1 diabetes. *Proc. Natl Acad. Sci. USA* **115**, E263–E272 (2018).
139. Manzoli, V. et al. Immunoisolation of murine islet allografts in vascularized sites through conformal coating with polyethylene glycol. *Am. J. Transplant.* **18**, 590–603 (2018).
140. Chen, T. et al. Alginate encapsulant incorporating CXCL12 supports long-term allo- and xenoislet transplantation without systemic immune suppression. *Am. J. Transplant.* **15**, 618–627 (2015).
141. Shoda, L. K. et al. A comprehensive review of interventions in the NOD mouse and implications for translation. *Immunity* **23**, 115–126 (2005).
142. Lernmark, A. & Larsson, H. E. Immune therapy in type 1 diabetes mellitus. *Nat. Rev. Endocrinol.* **9**, 92–103 (2013).
143. Reed, J. C. & Herold, K. C. Thinking bedside at the bench: the NOD mouse model of T1DM. *Nat. Rev. Endocrinol.* **11**, 308–314 (2015).
144. Keenan, H. A. et al. Residual insulin production and pancreatic β-cell turnover after 50 years of diabetes: Joslin Medalist Study. *Diabetes* **59**, 2846–2853 (2010).
     This paper demonstrated the persistence of insulin-expressing cells in patients with long-term T1D.
145. Liu, E. H. et al. Pancreatic β cell function persists in many patients with chronic type 1 diabetes, but is not dramatically improved by prolonged immunosuppression and euglycaemia from a β cell allograft. *Diabetologia* **52**, 1369–1380 (2009).
146. Dorrell, C. et al. Human islets contain four distinct subtypes of β cells. *Nat. Commun.* **7**, 11756 (2016).
     Refs 146 and 147 suggested that islet β-cells are heterogeneous in their molecular and functional properties.
147. Bader, E. et al. Identification of proliferative and mature β-cells in the islets of Langerhans. *Nature* **535**, 430–434 (2016).
148. Wang, Y. J. et al. Single-cell mass cytometry analysis of the human endocrine pancreas. *Cell Metab.* **24**, 616–626 (2016).
149. Johnston, N. R. et al. β Cell hubs dictate pancreatic islet responses to glucose. *Cell Metab.* **24**, 389–401 (2016).

Acknowledgements We apologize that we were unable to cite many studies owing to space limitations. We thank past and present members of our laboratories and colleagues for their insights and contributions. Q.Z. and D.A.M. receive support from National Institute of Health (NIH) and Harvard Stem Cell Institute (HSCI), and D.A.M. from Howard Hughes Medical Institute (HHMI).

Author contributions Q.Z. and D.A.M. wrote and edited the manuscript. Q.Z. prepared the figures.

Competing interests D.A.M. is a founder of Semma Therapeutics Inc. Q.Z. declares no competing interests.

Additional information
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to Q.Z. and D.A.M.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

REVIEW

https://doi.org/10.1038/s41586-018-0076-4

Restoring vision

Botond Roska[1,2,3]* & José-Alain Sahel[4,5,6,7]*

Restoring vision to the blind by retinal repair has been a dream of medicine for centuries, and the first successful procedures have recently been performed. Although we are still far from the restoration of high-resolution vision, step-by-step developments are overcoming crucial bottlenecks in therapy development and have enabled the restoration of some visual function in patients with specific blindness-causing diseases. Here, we discuss the current state of vision restoration and the problems related to retinal repair. We describe new model systems and translational technologies, as well as the clinical conditions in which new methods may help to combat blindness.

Reversing or slowing loss of vision has been an aim of ophthalmology since this discipline emerged[1] (Fig. 1). Surgical replacement of the lens has restored visual acuity in patients with cataracts[2] and the control of vascular leakage can stabilize or improve vision in those with wet age-related macular degeneration[3]. These treatments have had major social and economic impacts[4–7]. Despite these successes, there are still an estimated 285 million vision-impaired and 39 million blind people around the world[8]. Many conditions that cause blindness in underdeveloped countries are treatable because they involve refractive errors, cataracts, infections, or nutritional states, and the main problem is the delivery of existing care. The delivery of care in developed countries is not a major issue, but there are still no treatments for most blinding diseases, including advanced glaucoma, atrophic macular degeneration, advanced diabetic retinopathy, myopia, and monogenic retinal degeneration.

Here, we discuss the current state of vision restoration in blinding retinal diseases and the general problems related to retinal repair. We describe new model systems and translational technologies that have opened up new opportunities and given momentum to combating blindness—with human-derived retinal organoids and gene therapy representing the key advances. Finally, we discuss specific clinical conditions in which these methods may help to restore vision or slow retinal degeneration in the next decade.

Where we are today

Lens replacement after cataract and the control of fluid leakage from retinal blood vessels in wet age-related macular degeneration are now part of standard clinical care[2,3]. Over the past decade, we have witnessed major new developments in medical technology that have led to vision restoration or the slowing of vision loss in blinding diseases originating from retinal cells. A key example is gene therapy for a form of Leber congenital amaurosis (LCA)[9–11] that produced significant improvements in performance in a visually guided task in young adult patients[11,12], as well as increased activation[13] and long-term structural plasticity[14] of the visual cortex. This demonstrated the safety of ocular gene therapy and the potential benefit to patients[15]. Initial results from other gene therapy trials, involving conditions such as the hereditary blinding disease choroideraemia, have also indicated possible benefits[16,17]. A second example is electric stimulation of the retina in adult patients with photoreceptor degeneration[18], which evokes visual percepts and in some cases has provided form vision[19,20]. These results imply that information can flow from the retina to higher visual areas in blind patients if the optic nerve is intact. A third example is the transplantation of retinal pigmented epithelial cells behind the retina, which has produced possible visual improvements or stable vision in a few patients with age-related macular degeneration[21,22] or Stargardt disease[21], a monogenic form of macular degeneration[23]. Controlled studies with larger groups of patients are needed to verify the visual benefits, but the safety of such transplants, when done properly[24], is notable.

In recent years, we have not only seen the emergence of new therapies but have also learned new lessons about the state of visual pathways during human visual development. In a healthy visual system, light-driven developmental events occur during the first years of life, the so-called critical period[25]. Whether restoration of retinal output in congenitally blind patients after the critical period could ever benefit their vision has been investigated. A set of patients with congenital cataracts that prevented the formation of a visual image on the retina from birth received lens replacement surgery at or after eight years of age. This resulted in the acquisition of pattern vision[26–28] and demonstrated that a lack of high-resolution vision during the first few years of life does not necessarily prevent the acquisition of useful vision after treatment.

The problem of retinal repair

Apart from these major developments in therapies and in our understanding, advances in clinical care to treat blinding diseases have perhaps been slow from the point of view of the patient. Why is it difficult to slow down the progression of human vision loss or to restore vision once it has been lost? There are five central problems.

First, without intervention, intrinsic regeneration of the mammalian retina is weak or entirely absent. This is in contrast to some other animals, such as fishes, in which the generation of new retinal cells and regeneration of portions of a damaged retina occur after injury even in adulthood[29].

Second, the retina is a biological computer made up of about 100 different cell types that form specific synaptic connections and reside in functionally distinct microcircuits[30,31] (Fig. 2). Retinal cell types differ in their physiology, morphology, and patterns of gene expression[32,33]. As a consequence of this cell-type diversity, many diseases of vision are broadly or narrowly cell-type-specific because the disease-causing gene is expressed in a subset of cell types or because vulnerability is cell-type specific. To reduce side effects, the ideal treatment should also be

---

[1]Institute of Molecular and Clinical Ophthalmology Basel, Basel, Switzerland. [2]Neural Circuit Laboratories, Friedrich Miescher Institute, Basel, Switzerland. [3]Department of Ophthalmology, University of Basel, Basel, Switzerland. [4]Department of Ophthalmology, The University of Pittsburgh School of Medicine, Pittsburgh, PA, USA. [5]Institut de la Vision, Sorbonne Université, INSERM, CNRS, Paris, France. [6]CHNO des Quinze-Vingts, CIC INSERM-DGOS 1423, DHU Sight Restore, Paris, France. [7]Department of Ophthalmology, Fondation Ophtalmologique Rothschild, Paris, France. *e-mail: botond.roska@iob.ch; j.sahel@gmail.com

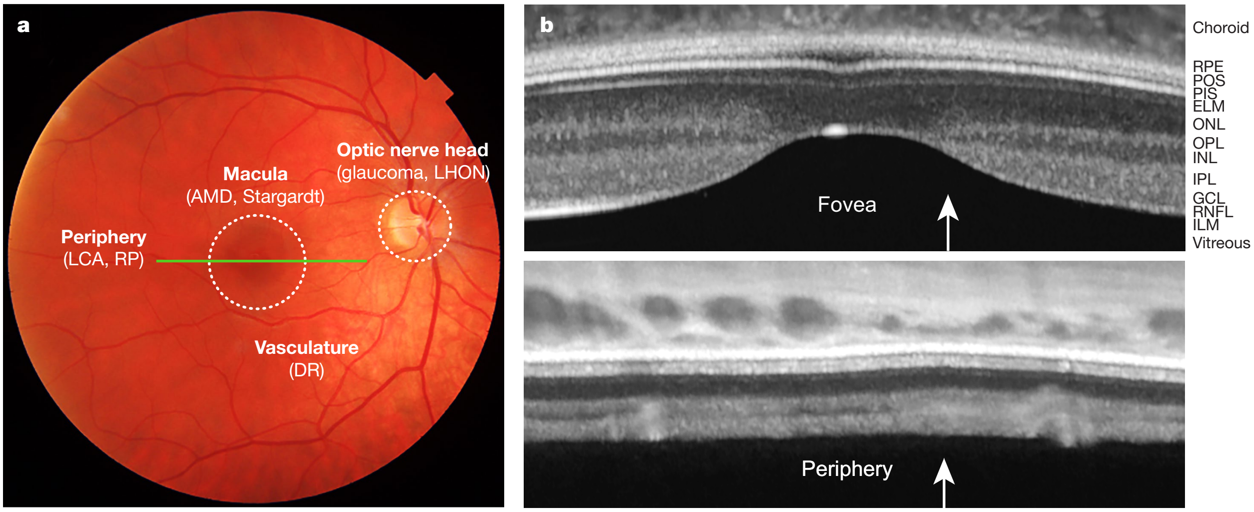


Fig. 1 | The human retina in vivo. a, The retina as seen at the back of eye by an ophthalmologist using an ophthalmoscope. Four objects can be detected: the vasculature, where diabetic retinopathy (DR) arises; the macula, the centre of which is called the fovea (age-related macular degeneration (AMD) and Stargardt disease affect this area); the optic nerve head, which is affected by glaucoma and Leber hereditary optic neuropathy (LHON); and the retinal periphery, which denotes the entire retina outside the macula, where Leber congenital amaurosis (LCA) and retinitis pigmentosa (RP) start. **b**, Cross sections of the retina (along the line shown in green in **a**) visualized using optical coherence tomography (OCT). Top: area of the macula. The fovea is marked by thinning of the retina. Bottom: retinal periphery. Different layers can be identified in an OCT image around the fovea. In general, cell body-rich layers and layers filled with liquid are darker. Layers from the vitreous towards the choroid (the layer behind the retina that is filled with capillaries) are: internal limiting membrane (ILM, the end feet of Müller cells on the vitreal side), retinal nerve fibre layer (RNFL, ganglion cell axons), ganglion cell layer (GCL, cell bodies of ganglion cells), inner plexiform layer (IPL, connections between ganglion cells and other cells in the inner retina, such as amacrine and bipolar cells), inner nuclear layer (INL, cell bodies of inner retinal neurons such as amacrine and bipolar cells), outer plexiform layer (OPL, connections between bipolar cells and photoreceptors), outer nuclear layer (ONL, photoreceptor cell bodies), external limiting membrane (ELM, the end feet of Müller cells on the choroidal side), photoreceptor inner segments rich in mitochondria (PIS), photoreceptor outer segments (POS), retinal pigmented epithelial cells (RPE), choroid. Distinct but fewer layers can also be distinguished in an OCT image of the retinal periphery, where the retina is thinner. White arrows show the direction of light.

cell-type specific, but it is difficult to achieve such treatment even in non-human models. A second consequence of cell-type diversity is that both the ability of gene therapy vectors to enter cells and the efficiency of transcription from vectors are cell-type-specific. Thus, vectors need to be tailor-made for different cell types[34]. Furthermore, specific synaptic connectivity in the retina has consequences for cell therapy. Repair of diseased retinas by introducing additional cells of a given type requires that the added cells connect to their natural partners, which is not efficient in adults[35].

Third, it is often assumed that findings in mice, the most common mammalian disease model, can be directly translated to humans. Although the overall cellular architecture of the retina is similar in mice and humans, the two species exhibit certain differences[36] that can complicate translation. There are differences in retinal cell-type composition, in cell-type-specific gene expression, and in the organization of the retina at the macroscopic level. For example, the human retina contains midget ganglion cells and their related retinal circuitry[37], but the mouse retina apparently does not. Midget cells are thought to mediate high-resolution image-forming vision[38], and preventing the degeneration of midget cells in patients with glaucoma or diabetic retinopathy would have a major effect on retarding disease progression—but there is no mouse model. Usher I genes are examples of differences in cell-type-specific gene expression between mice and humans. Mutant Usher I genes lead to Usher I syndrome, an inherited form of blindness paired with deafness. Mouse photoreceptors express Usher I proteins at much lower levels[39,40] than human photoreceptors, and lack the cellular compartment in which the proteins are localized in human photoreceptors[39]. Not surprisingly, mice bearing the same mutations as patients with Usher I syndrome have normal vision[39]. Finally, a striking difference in the organization of the retina at the macroscopic level between mice and humans is the presence in humans of a small compartment at the centre of the retina called the fovea. This is absent in mice[41]; in fact, primates are the only mammals with a fovea. The human fovea makes up only 0.2% of the retina[42] but is necessary for high-resolution colour vision[38], and is essential for important everyday tasks such as recognizing faces and reading[43]. Age-related macular degeneration[3] and Stargardt disease[23] both initially affect the fovea.

Fourth, the eyes of primates have a thick membrane, the so-called inner limiting membrane, between the retina and the inner compartment of the eye (the vitreous). This membrane limits diffusion and thus restricts the efficacy of intravitreally delivered gene therapy vectors, such as adeno-associated viral vectors (AAVs)[44]. Newly constructed AAVs show slightly increased peripheral expression, but access is still limited[45]. However, the limitation of diffusion across the inner limiting membrane is lower in the foveal region of the retina and gene transfer by intravitreal AAV injections around the fovea in primates has been successful[46]. The difficulty in transducing the peripheral retina with gene therapy vectors is important, as intravitreal injections are easier for ophthalmologists and safer to perform than sub-retinal injections[45].

Fifth, the surface area of the human retina is very large (about 1,000 mm^2). It is seventy times larger than that of the mouse (about 15 mm^2)[47,48]. In contrast to treatments with small molecules, which distribute across the retina when injected into the eye or delivered via the bloodstream, scaling up gene delivery to reach all cells of a given type across the human retina is a daunting task.

New model systems

Given these difficulties related to retinal repair and the differences between the retinas of humans and other species, how can we understand human retinal diseases and develop new therapies? Historically, researchers have used genetically tractable small animal models and cell lines to study retinal diseases and to test treatment strategies. While work on these systems will continue, three new retina models have become available in recent years that may increase our understanding of human retinal disease and benefit the search for new therapies (Table 1).

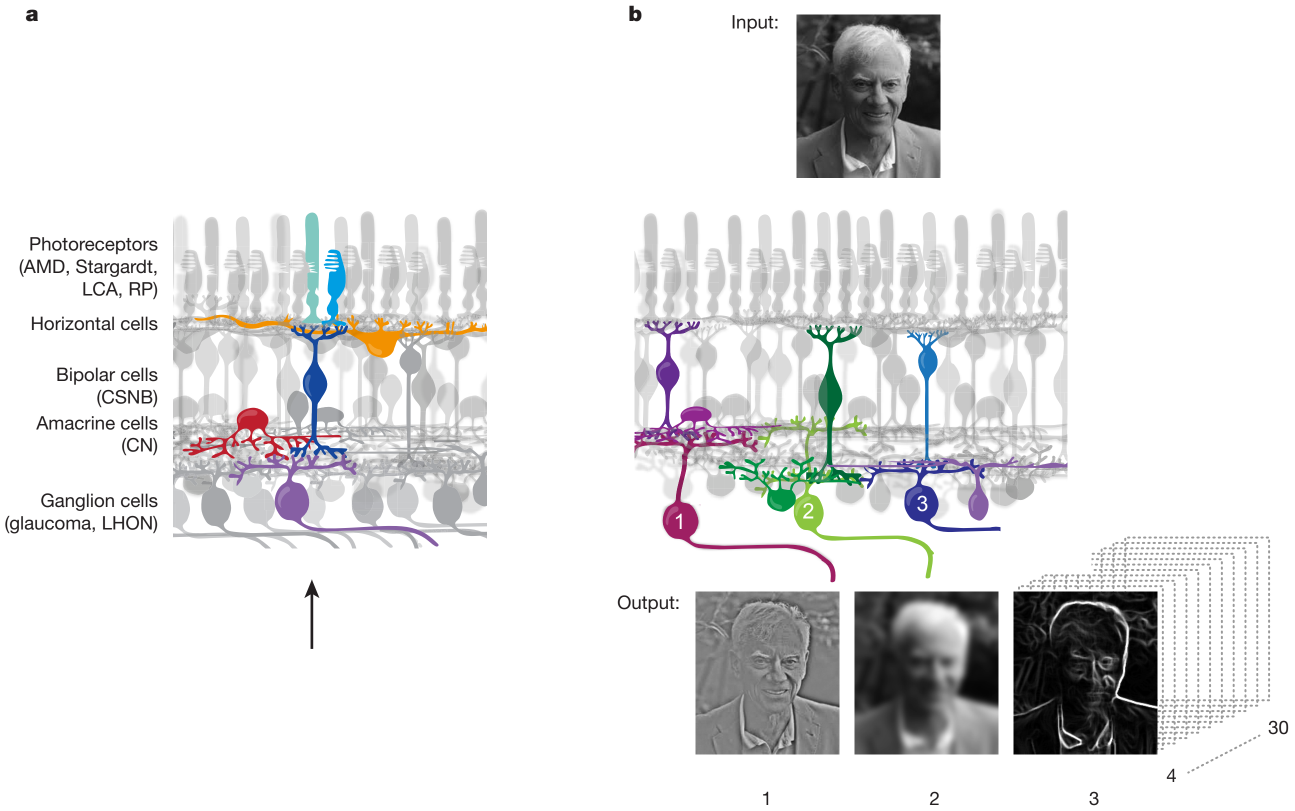


Fig. 2 | Cell types, circuits and computations performed by the vertebrate retina. a, Retinal cell types (about 100 types) can be divided into a few major classes. The characteristic locations of the cell bodies, gross morphology and connectivity of some of these classes in the retinal periphery are shown. Photoreceptors (rods in cyan, and cones in light blue) capture light. Photoreceptors pass information to inhibitory horizontal cells (orange) and excitatory bipolar cells (dark blue). Horizontal cells feed back to photoreceptors, while bipolar cells pass information further to inhibitory amacrine cells (red) and excitatory ganglion cells (purple). Amacrine cells either feed back to bipolar cells or pass information to ganglion cells. Ganglion cells are the output neurons of the retina, sending information to the rest of the brain via their axons, which come together to form the optic nerve. This circuit layout is highly conserved across vertebrates. The retinal circuit is similar in the fovea, but the locations of the cell bodies are different: cell bodies of all cell classes except cones are displaced to the side, which allows light to hit the cones directly. The black arrow shows the direction of light. Different diseases (in parentheses) affect different cell classes. AMD, age-related macular degeneration; LCA, Leber congenital amaurosis; RP, retinitis pigmentosa; CSNB, congenital stationary night blindness; CN, congenital nystagmus; LHON, Leber hereditary optic neuropathy. **b**, Retinal cell types are organized into about 30 circuits, each of which receives input from the mosaic of photoreceptors and ends with a mosaic of ganglion cells of a given type that forms the retinal output. There are about 30 types of ganglion cell and the retina therefore creates about 30 different image representations (three examples are shown at the bottom) of the original image that enters the eye (the face at the top). The retina is thus a parallel image processor that describes to the brain the image that falls on the retina using its own 'language'.

Human retinal organoids

A simple skin biopsy from a patient with a hereditary retinal disease can be induced within one year to generate layered retina-like neuronal structures, termed retinal organoids. These consist of different cell types and appear to bear genetic markings similar to those of retinal neurons[49–51]. Alternatively, retinal organoids grown from a healthy human biopsy can be engineered to carry relevant mutations[52]. Such organoids could enable unprecedented insights into the development of retinal diseases and provide a controlled, biologically realistic system in which to test new treatment strategies. Disease-model retinal organoids could prove useful in several ways. With diseases that cause symptoms early in life, a disease phenotype may develop in the organoids, thus allowing investigation of the mechanistic basis of the disease and the development of repair strategies. The extent to which retinal organoids can develop into adult tissue[53] is currently unclear and thus disease-model retinal organoids for hereditary diseases in which patients develop symptoms later in life may not present relevant phenotypes. However, such retinal organoids could still be useful for testing cell-type-specific gene delivery or the efficacy of gene editing in selected cell types. Finally, organoid culture allows control of the growth environment of human retina-like tissue and can thus be used for experiments previously impossible with human retinas, such as the precise investigation of physical and chemical stresses (for example, pressure, levels of oxygen or glucose) that may underlie disease progression.

Table 1 | The advantages and limitations of different model systems

	Zebrafish	Mouse	Marmoset	Human	Organoid
Fovea	No	No	Yes	Yes	No
In vivo	Yes	Yes	Yes	No	No
In vitro	Yes	Yes	Yes	Yes	Yes
In vivo screen	Yes	No	No	No	No
In vitro screen	No	No	No	No	Yes
Adult	Yes	Yes	Yes	Yes	No
Genetic engineering	Yes	Yes	Yes	No	Yes
Availability	+++	+++	+	++	+++
Cost	+	++	+++	+	++

Research combined in a variety of model systems allows a better understanding of the pathomechanism of disease and the development of therapy.

Post-mortem human retinas

Retinas from human donors can be kept alive in culture for at least three weeks post mortem[54,55], and recent improvements in culture conditions may substantially extend their longevity. This is a key model system for understanding the cell types and circuits of the adult human retina. The recording of functional light responses at cellular resolution from the retinas of animal models in vitro has been possible for years[56] and

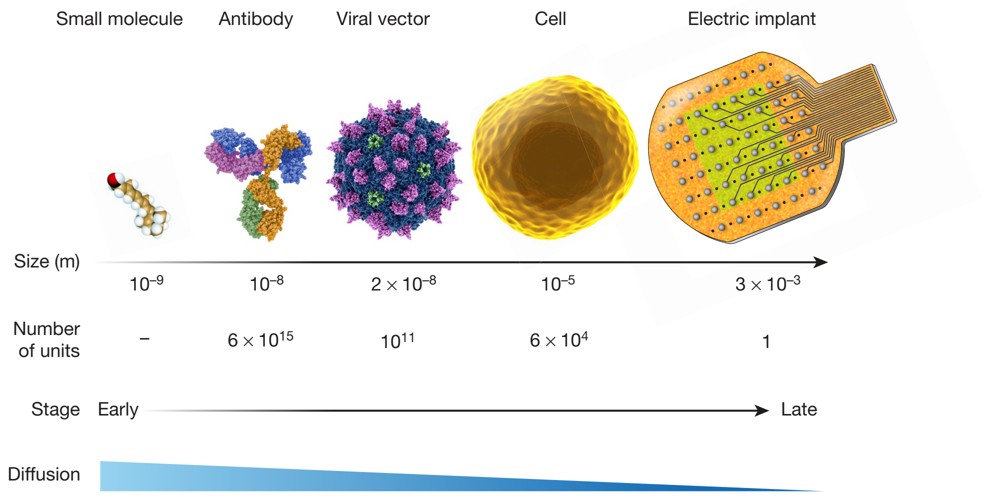


Fig. 3 | Current and planned tools for treating blinding diseases. The approximate physical sizes span six log units, and the numbers of tools delivered to the eye span 15 log units. The approximate sizes and numbers of tools delivered are shown below each tool. The complexity and sophistication of the tools increase with size, as does the difficulty of delivering treatment across a large retinal surface in vivo. The characteristics of the different tools are shown below the images.

further technical advances will allow this to also be done in human retinas. This will improve the comparison of retinal responses with visual behaviour in humans and facilitate direct correlation of gene expression and physiological responses for human retinal cell types. Furthermore, direct access to human retinas will allow the testing of gene therapy vectors, cell transplantation, and neuroprotection strategies in a more controlled manner and with a higher throughput than is currently possible: a single human retina can be dissected into many smaller pieces that can be examined individually.

Genetic disease models in marmosets

Non-human primates have been used for a number of years to study vision and vision loss, with most of the classic work done in macaques. However, it is difficult to apply transgenic methods to this species and the marmoset has emerged recently as a much easier non-human primate model for transgenic work[57]. Importantly, marmoset retinas have a fovea[58], making the marmoset more relevant for developing treatment strategies than non-primate (that is, non-foveated) models such as mice. Thus, the advancement of transgenic applications in the marmoset as an in vivo, foveated model of human visual diseases is of crucial importance for future vision research.

New translational technologies

In order to treat blindness, we need not only new model systems but also new translational technologies. While treatments based on small molecules, antibodies, and surgery remain the main therapeutic approaches used today, new technologies for repairing damaged retinas and restoring light sensitivity to impaired retinas are also being developed (Fig. 3). We will discuss four new technologies here.

Gene therapy

Gene therapy[59] has gained momentum with the introduction of AAVs[60] as vectors for delivering genes of interest to retinal cells[61,62]. AAVs diffuse easily in tissue, have little toxicity, and bring about sustained gene expression, although their integration into the human genome is minimal[63]. AAVs have three important components: the capsid for cell entry[64], the promoter driving transgene expression and cell-type-specific expression[62,65], and the therapeutic gene as the viral payload. A weakness of AAVs is their relatively small carrying capacity, which limits their application to small genes. As the efficacies of both capsid and promoter depend on cell type and species, AAVs optimized in mice are unlikely to be efficient and cell-type-specific in humans. Screening of AAVs in vitro in post mortem human retinas[54,55] and in vivo in primates eyes[65] could identify vectors that are relevant for humans. Post mortem retinas can be used to identify the cell types that are targeted in adult human retinas; primate retinas can reveal the efficacy of infection in living foveated eyes. However, given an effective AAV for delivering a gene of interest to a desired cell type in the human retina, the challenge remains of reliably and safely delivering AAVs to the entire human retina. As mentioned above, shielding of the human retina by the inner limiting membrane restricts infection to the foveal area. Sub-retinal injections may provide access to a larger retinal area but this is technically more complicated and separates the retina from the retinal pigmented epithelium, with the danger of photoreceptor damage[66]. Furthermore, the human retina is so large that a single AAV injection site is likely to be insufficient for full coverage. Despite these limitations, gene therapy via AAVs is likely to remain a leading strategy for the treatment of vision loss in the coming years. Non-viral gene delivery approaches[67–69] are also being pursued, such as laser, ultrasound, and electrical discharges, which create momentary pores in cell membranes and allow DNA or RNA fragments to enter the cell. Unlike AAV-based delivery, such non-viral gene delivery technologies produce transient expression of the target protein.

Cell therapy

Cell therapy provides retinal repair via ectopic cell transplantation to replace damaged or dead cells. Cells can be obtained from cell lines derived from embryonic or induced pluripotent stem cells, or alternatively from retinal organoids in which the developmental stage of the transplanted cells can be controlled. Cell therapy via retinal pigmented epithelial cells[21] appears to be simpler than via retinal neurons such as photoreceptors, because pigmented epithelial cells do not need to form neuronal signal-carrying contacts such as synapses. Nevertheless, transplanted epithelial cells need to interact with photoreceptors to be useful for retinal function. Cell therapy for retinal neurons is more complex because transplanted cells need to migrate to the correct location, make the proper synaptic connections, and exhibit cell-type-specific physiological responses. While there has been some success in transplanting photoreceptors into mouse retinas[70,71], it is not clear that current strategies will be efficient and effective in a clinical setting[72]. Only a small proportion of transplanted cells appears to survive in mice and make the necessary synaptic connections[71], and the transfer of material between transplanted and native photoreceptors complicates interpretation of the results[73–75]. Recently, transplanted ganglion cells were shown to integrate into retinal circuits, providing yet another cellular source for repair[76,77]; however, it has not yet been shown that the transplanted ganglion cells grow axons that reach their targets in the brain. Future experiments with retinal pigmented epithelial cells, photoreceptors and ganglion cells, in both animal models and in patients, will further define the feasibility of this therapeutic approach.

Induced retinal regeneration

Induced retinal regeneration attempts to apply lessons learned from animals, such as fishes, in which regeneration of the injured retina occurs. In response to damage to the retina, Müller glial cells in fish return to a stem-cell-like state, divide, and develop into different cell types that integrate into the retinal circuitry[29]. Mammals do not have this innate capability, but research on experimentally driven regeneration is in progress. For example, mouse Müller cells activated by specific small molecules or genes start to divide[78] and the newly generated cells integrate into the retinal circuitry[79]. Induced retinal regeneration has also been used to induce the regrowth of ganglion cell axons towards their brain targets[80,81]. The cut axons of retinal ganglion cells regrow axonal projections into the brain when provided with appropriate genes, small molecules, or neuronal activity[81–83]. The axons even reach their correct cell-type-specific targets and establish functional connections[81]. While these methods are currently performed only in animal models and their efficiency is relatively low, future work is likely to improve performance and add induced retinal regeneration as an important tool to the translational repertoire.

Artificial retinal stimulation

Artificial retinal stimulation using electronic implants, optogenetics or photoswitches is designed to restore the image-acquiring properties of the retina[84,85]. The patient wears a camera mounted, for example, on a pair of spectacles. The video signal is converted into an electronic pattern or a pattern of light, depending on the vision restoration strategy. In electronic implant therapy, the implant is placed in physical contact with the retina and an electronic pattern created by implant electrodes stimulates the remaining light-insensitive retinal cells. The current generation of retinal implants provides relatively modest visual improvement, owing partly to the low density and low number of electrodes in the implants[20,86], as well as to the large distances between implanted electrodes and target cells[87]. Cortical implants are an alternative to retinal implants and a cortical implant that provides surface stimulation to the primary visual cortex is in clinical trials[88]. In optogenetic therapy, monochromatic light activates light-gated ion channels delivered by gene therapy to remaining specific cell types in the blind retina[54,89,90]. In photoswitch therapy[91], monochromatic light activates light-sensitive small molecules delivered to retinal cells that couple to intrinsic channels on the surface of retinal cells. Theoretically, these latter approaches act at a higher resolution than electronic implants and can activate specific retinal cell types, giving increased control of specific retinal microcircuits. Limitations of current optogenetic and photoswitch technologies are that the activation of cells requires bright visible light and that the dynamic range of 2–3 logarithmic units is narrow[92]. In a number of blinding diseases, including macular degeneration, most visual sensitivity is lost in a specific region of the retina, while other regions remain light sensitive. The remaining light-sensitive retinal regions rule out or impede the use of bright visible light for activation. Although it is possible to increase the sensitivity of optogenetic sensors, this often slows the response[93], and thus rules out restoration of useful motion vision. Furthermore, since the dynamic range remains narrow, this approach still requires an external camera. The development of near-infrared optogenetic sensors or photoswitches could mitigate the problem of the remaining light sensitivity. Optogenetic approaches are currently in early clinical development and higher resolution electronic implants are being developed. In the coming years, different forms of artificial retinal stimulation are likely to be important strategies in vision-loss therapy.

Implementing technologies for therapy

The various translational technologies outlined above could be used either alone or in combination to treat a variety of diseases. However, the decision about which technology is suitable for which disease and at which stage requires a detailed understanding of the different blinding diseases and their natural history. For example, it is important to discriminate loss of function, which is potentially reversible, from loss of cells, which is irreversible. In general, the therapeutic strategy depends on the stage of the disease. If vision has severely deteriorated or is completely lost, the aim has to be to restore vision; if significant vision is still present, the aim may be to prevent or slow further loss. As is typical in medicine, the border between these stages is not sharp but the two ends of the spectrum do allow us to highlight differences in the ways in which the available technologies can be used.

Vision is lost

If the patient has lost all useful vision, the only therapeutic option is vision restoration. The available technological repertoire is wide, since only general safety considerations are important, such as avoidance of an immune response or the malignant transformation of cells. Importantly, restoring vision even in a small patch of the retina may lead to useful vision and, thus, the new technologies listed above that lead to regained light sensitivity would be promising approaches. Most vision restoration strategies that are considered today in clinical settings aim to target diseases stemming from photoreceptor degeneration.

Cones are the photoreceptors that are responsible for high-resolution colour vision under daylight conditions, and rods are responsible for vision under dim light conditions. Many monogenic diseases primarily affect rods. Late-onset forms of these monogenic diseases are called retinitis pigmentosa[94]; Leber congenital amaurosis[95] describes the condition in which symptoms of retinitis pigmentosa are present during infancy. Although rods in humans outnumber cones 20 to 1, we artificially illuminate our environment today to such an extent that individuals without functional rods can lead their lives relatively unaffected. However, for reasons not yet fully understood, rod degeneration leads to secondary cone degeneration, even when the gene associated with the disease is not expressed in cones[59]. This means that the most severe effect of rod loss is the subsequent loss of cones and cone-derived vision. Repair strategies at later stages of retinitis pigmentosa aim to connect the retina to the visual world by providing artificial retinal stimulation using electronic implants, optogenetics, or photoswitches. An alternative is the transplantation of photoreceptors[59,84,85].

Vision restoration strategies are also being considered that target diseases in which ganglion cells degenerate. After ganglion cells have died, such as in advanced stages of glaucoma or diabetic retinopathy, there is currently little that can be done. But there are at least three ways to develop potential therapy. First, to restore vision at higher levels in the visual system such as the thalamus or visual cortex via electronic implants[96] or optogenetic stimulation. Second, to transplant ganglion cells[77]. Third, to perform whole eye[97] or organoid transplantation. With the exception of a cortical implant now in clinical trials[88], these approaches are in their early stages or are simply theoretical options with no proof of concept as yet.

Useful vision remains

If the patient still has some useful vision, the best approach for curing blindness is the prevention, or at least slowing, of vision loss. Here, a major factor when designing a new therapy is to guarantee that the treatment does not interfere in any way with the remaining natural vision. One key point is to protect foveal vision. A second factor concerns the distribution of the therapeutic agents. If a blinding disease is progressive and affects the entire retina, the molecules or genes to slow visual loss need to be delivered to the disease-affected retinal cell types across the whole retina. With the exception of the fovea, delivering therapeutic agents to a small patch of the retina will be likely to have little effect on overall degeneration and loss of vision. Therefore, approaches to slow vision loss differ between fovea-affecting diseases and diseases that affect the whole retina. Since we are far from being able to deliver genes across the whole human retina, retina-wide delivery is most likely to succeed for small-molecule or antibody-based treatments, or by using local gene therapy that produces secreted factors that spread over a wide area.

A successful example of slowing down vision loss is the intravitreal injection of anti-VEGF antibodies in wet age-related macular degeneration; this decreases the leakage of fluid from abnormal retinal blood vessels and its destructive accumulation (oedema). However, while this approach substantially delays vision loss, anti-VEGF therapies target only the vascular component of the disease and mostly reduce blood leakage. The causative and amplifying processes (for example, genetic variants, inflammatory and immune cell activation and photoreceptor damage) are not treated[98]. In atrophic age-related macular degeneration, cell therapy involving transplantation of retinal pigmented epithelial cells is the approach that is gaining the most traction[99]. Larger clinical trials will show whether this approach increases vision in patients.

Our most precious photoreceptors are the foveal cones, the loss of which leads to severe visual handicap. Juvenile macular degeneration, including Stargardt disease, and atrophic age-related macular degeneration mostly affect foveal cones. Although Stargardt disease is a target for gene therapy[23], the most frequently mutated gene in this disease, *ABCA4*, is too large to fit into an AAV vector. The options to be considered here are the use of alternative viral vectors, trimming the length of the gene, or its assembly from shorter pieces within target cells. Gene therapy using *ABCA4* in larger-capacity lentiviral vectors is now in clinical trials[100].

Recent promising developments have reduced secondary degeneration of cones in animal models of retinitis pigmentosa. The delivery of rod-derived cone viability factor[101–103] or activation of its receptor[104], or the delivery of genes countering energy starvation[105,106] or oxidative stress[107], have been shown to decrease cone degeneration in animal models and are being considered for human therapy. Starvation[108,109] and oxidative stress[110,111] appear to be major contributors to cone degeneration after rod death; therefore, feeding cones and preventing oxidative damage emerge as central themes in their protection. Methods that may increase nutrient availability by increasing blood supply[112], such as microcurrent stimulation[113,114], are in clinical trials, while methods to reduce oxidative stress via antioxidant supplementation are already in clinical practice[115].

There have also been developments in the reduction of vision loss in a form of Leber hereditary optic neuropathy (LHON)[116]. Mutations in different genes encoded by the mitochondrial genome cause different forms of LHON. This disease is a clear example of cell-type-specific vulnerability; despite the presence of mitochondria in every cell in the body, only or mostly retinal ganglion cells degenerate. Gene therapy and small molecule-based therapies have both shown promising slowing of vision loss in LHON.

Everything we have described so far relates to retinal vision loss; however, in a minority of cases vision loss is due not to retinal problems but rather to defects in other regions of the visual system, such as the visual cortex. In these conditions, influencing the plasticity of the remaining visual cortex has been invoked as a way to improve visual function[117].

Evaluating the outcome of therapy

The emergence of new therapies to slow loss of vision or to restore vision creates the need for major innovation and international collaboration on ways to measure the outcome of therapies in patients.

One key challenge is variability in the results of clinical trials. For example, one group reported increased visual function[12] after LCA gene therapy, while other groups found initial improvement that declined over time[118,119], or even an overall decline in visual function compared to baseline in some patients[118]. Variability has also been observed in the visual benefit to patients of electronic implants. Variability can originate in many ways: different ways of evaluating visual function, measuring visual function at different times after treatment, differences in the details of gene therapy vectors and their delivery[66], the details of how visual stimulation devices are implanted, the ability of patients to learn the meaning of altered peripheral input, and differences in the disease stages at which the therapy was provided.

Another challenge is that retinal degenerative diseases are progressive and it is therefore essential to study progression of the disease before enrolling patients in clinical trials, and to characterize changes in progression during and following therapy[85]. Moreover, new and existing diagnostic tools should be incorporated into the development of a standardized and quantitative evaluation scheme to assess retinal and cortical[117] structure and function as well as visual function in both control patients and patients receiving therapy.

New and improved in vivo imaging methods (including adaptive optics, optical coherence tomography, ultrasound and autofluorescence imaging) and psychophysical methods (for example, microperimetry) can assist such evaluations and comparisons by allowing determination of the density of photoreceptors, the existence and length of inner and outer segments, the status of ganglion cells, interactions between photoreceptors and retinal pigmented epithelial cells, retinal blood flow[112], and any remaining function in a small patch of the retina[85]. However, demonstration of a clinically relevant benefit extends beyond such quantitative measures. Regulatory agencies and funding bodies will require real-life performance-based tests, such as mobility testing[15], as well as patient-reported outcomes of treatment. Measures of outcome must be robust, reproducible, relevant and easy to duplicate across various centres[120–122].

Rehabilitation is another key issue, since the outcome of therapy can improve with time and a well-designed rehabilitation program[123]. Improvement may be gradual because the connectivity of the visual cortex may have reconfigured during the state of blindness to support other cognitive functions[124]. Therefore, it may require visual training and time to readjust the visual cortex to its original function. Furthermore, vision restoration approaches in which technologies converge will require new rehabilitation strategies. As an example, optogenetic therapy combines gene therapy with a medical device, a goggle. This technology convergence will require different types of expertise and strong patient engagement during rehabilitation.

Technologies that complement vision restoration

Aside from strategies that explicitly set out to restore vision or slow vision loss, there are many alternative approaches currently available or being developed for conveying visual information to visually impaired patients through one of their remaining sensory modalities. Classic strategies include the use of guide dogs and canes. However, the digital revolution is also adding to the capabilities of blind patients: smartphone applications provide services as varied as using the phone as a simple light detector, or as a GPS-linked talking map, or to perform voice-written emailing, word processing and web browsing[125,126]. It has recently become possible to describe particular images to blind people using artificial intelligence[127]. As the power of artificial intelligence increases, more and more digital tools that are useful for blind individuals will be generated.

Other emerging technologies focus on the use of more complex sensory substitution, where the remaining senses are trained to process information normally processed by the visual system. One approach that has brought the highest recorded 'visual' acuity for any vision rehabilitation strategy applied to vision-impaired patients is visual-to-auditory sensory substitution[128]. Here a head-mounted camera records the visual environment and the video signal is then converted into a series of sounds. Remarkably, visual-to-auditory sensory substitution activates 'visual' brain areas with auditory stimulation[129]. Another approach is the use in humans of echolocation similar to that of bats and dolphins[130]. Although these advances may make life easier for the blind, the goal of medicine remains the prevention of vision loss in the first place and the cure of vision loss that has already occurred.

Outlook

Blinding diseases of the eye, which are increasing in step with increasing human longevity, cause major social and economic burdens[131]. The recent developments outlined in this paper provide the hope that some of these diseases can be effectively addressed in the coming decade. Nevertheless, the challenge is severe because the retina is a biological computer and not just a simple collection of photosensitive cells. Therefore, we argue that both the diagnosis and the treatment of retinal diseases should be cell-type focused: disease by disease, we need to understand the causes and mechanisms of the death of disease-affected human retinal cell types. We must also learn to interact with retinal cell types inside the patient's eye, specifically and efficiently.

Finally, the retina, as a biological computer comprising about 100 cell types with specific rules of connectivity, is not unique. The cerebral cortex[32] and some deep brain nuclei[132] are both cell-type rich and have specific connectivity. The retina is in fact an outpost of the brain sitting in the eye[56] and following rules similar to the rest of the brain. This is important for two reasons. First, key pathological phenomena, including cell-type specific degeneration, are shared by the retina and other brain regions. Understanding why retinal cells die and why this process is cell-type specific could enhance our understanding of neurodegeneration in general. Second, fine perturbations of the activity of specific retinal cell types are likely to occur in a variety of neurological and psychiatric diseases with a genetic or environmental origin, because the disease-associated genes are expressed in specific retinal cell types or because the environmental insult affects specific retinal cell types. The fact that the retina is visible through the lens in humans makes it the part of the brain that is simplest to study in patients in vivo. The development of new imaging techniques to record the activity of retinal cells at cellular resolution in vivo in humans would make the retina the ideal location for following the progress of brain disease and reaction to therapy.

Received: 1 October 2017; Accepted: 16 February 2018;
Published online 16 May 2018.

1. The Lasker Foundation. *Restoring Vision to the Blind* http://www.laskerfoundation.org/new-noteworthy/articles/restoring-vision-blind/ (2015).
2. Olson, R. J. Cataract surgery from 1918 to the present and future—just imagine! *Am. J. Ophthalmol.* **185**, 10–13 (2017).
3. Jager, R. D., Mieler, W. F. & Miller, J. W. Age-related macular degeneration. *N. Engl. J. Med.* **358**, 2606–2617 (2008).
4. Hodgson, N. et al. Economic and quality of life benefits of anti-VEGF therapy. *Mol. Pharm.* **13**, 2877–2880 (2016).
5. Finger, R. P., Guymer, R. H., Gillies, M. C. & Keeffe, J. E. The impact of anti-vascular endothelial growth factor treatment on quality of life in neovascular age-related macular degeneration. *Ophthalmology* **121**, 1246–1251 (2014).
6. Essue, B. M. et al. A multicenter prospective cohort study of quality of life and economic outcomes after cataract surgery in Vietnam: the VISIONARY study. *Ophthalmology* **121**, 2138–2146 (2014).
7. Lamoureux, E. L., Fenwick, E., Pesudovs, K. & Tan, D. The impact of cataract surgery on quality of life. *Curr. Opin. Ophthalmol.* **22**, 19–27 (2011).
8. WHO. *Visual Impairment and Blindness* http://www.who.int/mediacentre/factsheets/fs282/en/ (2017).
9. Bainbridge, J. W. B. et al. Effect of gene therapy on visual function in Leber's congenital amaurosis. *N. Engl. J. Med.* **358**, 2231–2239 (2008).
10. Hauswirth, W. W. et al. Treatment of Leber congenital amaurosis due to RPE65 mutations by ocular subretinal injection of adeno-associated virus gene vector: short-term results of a phase I trial. *Hum. Gene Ther.* **19**, 979–990 (2008).
11. Maguire, A. M. et al. Safety and efficacy of gene transfer for Leber's congenital amaurosis. *N. Engl. J. Med.* **358**, 2240–2248 (2008).
This study led to the first FDA-approved AAV gene therapy.
12. Russell, S. et al. Efficacy and safety of voretigene neparvovec (AAV2-hRPE65v2) in patients with RPE65-mediated inherited retinal dystrophy: a randomised, controlled, open-label, phase 3 trial. *Lancet* **390**, 849–860 (2017).
The first FDA-approved AAV gene therapy.
13. Ashtari, M. et al. The human visual cortex responds to gene therapy-mediated recovery of retinal function. *J. Clin. Invest.* **121**, 2160–2168 (2011).
14. Ashtari, M. et al. Plasticity of the human visual system after retinal gene therapy in patients with Leber's congenital amaurosis. *Sci. Transl. Med.* **7**, 296ra110 (2015).
15. Ledford, H. FDA advisers back gene therapy for rare form of blindness. *Nature* **550**, 314 (2017).
16. MacLaren, R. E. et al. Retinal gene therapy in patients with choroideremia: initial findings from a phase 1/2 clinical trial. *Lancet* **383**, 1129–1137 (2014).
17. Edwards, T. L. et al. Visual acuity after retinal gene therapy for choroideremia. *N. Engl. J. Med.* **374**, 1996–1998 (2016).
18. Humayun, M. S. et al. Visual perception elicited by electrical stimulation of retina in blind humans. *Arch. Ophthalmol.* **114**, 40–46 (1996).
This work led to the development of epiretinal implants.
19. Mills, J. O., Jalil, A. & Stanga, P. E. Electronic retinal implants and artificial vision: journey and present. *Eye* **31**, 1383–1398 (2017).
20. da Cruz, L. et al. Five-year safety and performance results from the Argus II retinal prosthesis system clinical trial. *Ophthalmology* **123**, 2248–2254 (2016).
21. Schwartz, S. D. et al. Human embryonic stem cell-derived retinal pigment epithelium in patients with age-related macular degeneration and Stargardt's macular dystrophy: follow-up of two open-label phase 1/2 studies. *Lancet* **385**, 509–516 (2015).
22. Mandai, M. et al. Autologous induced stem-cell-derived retinal cells for macular degeneration. *N. Engl. J. Med.* **376**, 1038–1046 (2017).
23. Tanna, P., Strauss, R. W., Fujinami, K. & Michaelides, M. Stargardt disease: clinical features, molecular genetics, animal models and therapeutic options. *Br. J. Ophthalmol.* **101**, 25–30 (2017).
24. Daley, G. Q. Polar extremes in the clinical use of stem cells. *N. Engl. J. Med.* **376**, 1075–1077 (2017).
25. Espinosa, J. S. & Stryker, M. P. Development and plasticity of the primary visual cortex. *Neuron* **75**, 230–249 (2012).
26. Ganesh, S. et al. Results of late surgical intervention in children with early-onset bilateral cataracts. *Br. J. Ophthalmol.* **98**, 1424–1428 (2014).
27. Kalia, A. et al. Development of pattern vision following early and extended blindness. *Proc. Natl Acad. Sci. USA* **111**, 2035–2039 (2014).
28. Sinha, P., Chatterjee, G., Gandhi, T. & Kalia, A. Restoring vision through "Project Prakash": the opportunities for merging science and service. *PLoS Biol.* **11**, e1001741 (2013).
29. Wan, J. & Goldman, D. Retina regeneration in zebrafish. *Curr. Opin. Genet. Dev.* **40**, 41–47 (2016).
30. Azeredo da Silveira, R. & Roska, B. Cell types, circuits, computation. *Curr. Opin. Neurobiol.* **21**, 664–671 (2011).
31. Masland, R. H. The neuronal organization of the retina. *Neuron* **76**, 266–280 (2012).
32. Zeng, H. & Sanes, J. R. Neuronal cell-type classification: challenges, opportunities and the path forward. *Nat. Rev. Neurosci.* **18**, 530–546 (2017).
33. Seung, H. S. & Sümbül, U. Neuronal cell types and connectivity: lessons from the retina. *Neuron* **83**, 1262–1272 (2014).
34. Weinmann, J. & Grimm, D. Next-generation AAV vectors for clinical use: an ever-accelerating race. *Virus Genes* **53**, 707–713 (2017).
35. Santos-Ferreira, T. F., Borsch, O. & Ader, M. Rebuilding the missing part—a review on photoreceptor transplantation. *Front. Syst. Neurosci.* **10**, 105 (2017).
36. Chalupa, L. M. & Williams, R. W. (eds) *Eye, Retina, and Visual System of the Mouse* (MIT Press, Cambridge, 2008).
37. Dacey, D. M. The mosaic of midget ganglion cells in the human retina. *J. Neurosci.* **13**, 5334–5355 (1993).
38. Kolb, H. in *Webvision: The Organization of the Retina and Visual System* (eds Kolb, H. et al.) (Univ. Utah Health Sciences Center, Salt Lake City, 1995).
39. Sahly, I. et al. Localization of Usher 1 proteins to the photoreceptor calyceal processes, which are absent from mice. *J. Cell Biol.* **199**, 381–399 (2012).
Example of the limitations of mice as a disease model.
40. Siegert, S. et al. Transcriptional code and disease map for adult retinal cell types. *Nat. Neurosci.* **15**, 487–495 (2012).
41. Kostic, C. & Arsenijevic, Y. Animal modelling for inherited central vision loss. *J. Pathol.* **238**, 300–310 (2016).
42. Kolb, H., Fernandez, E. & Nelson, R. (eds) *Webvision: The Organization of the Retina and Visual System* (Univ. Utah Health Sciences Center, Salt Lake City, 1995).
43. Seiple, W., Rosen, R. B. & Garcia, P. M. T. Abnormal fixation in individuals with age-related macular degeneration when viewing an image of a face. *Optom. Vis. Sci.* **90**, 45–56 (2013).
44. Dalkara, D. et al. Inner limiting membrane barriers to AAV-mediated retinal transduction from the vitreous. *Mol. Ther.* **17**, 2096–2102 (2009).
45. Dalkara, D. et al. In vivo-directed evolution of a new adeno-associated virus for therapeutic outer retinal gene delivery from the vitreous. *Sci. Transl. Med.* **5**, 189ra76 (2013).
46. Yin, L. et al. Intravitreal injection of AAV2 transduces macaque inner retina. *Invest. Ophthalmol. Vis. Sci.* **52**, 2775–2783 (2011).
47. Panda-Jonas, S., Jonas, J. B., Jakobczyk, M. & Schneider, U. Retinal photoreceptor count, retinal surface area, and optic disc size in normal human eyes. *Ophthalmology* **101**, 519–523 (1994).
48. Remtulla, S. & Hallett, P. E. A schematic eye for the mouse, and comparisons with the rat. *Vision Res.* **25**, 21–31 (1985).
49. Eiraku, M. et al. Self-organizing optic-cup morphogenesis in three-dimensional culture. *Nature* **472**, 51–56 (2011).
This study led to the development of human retinal organoids.
50. Nakano, T. et al. Self-formation of optic cups and storable stratified neural retina from human ESCs. *Cell Stem Cell* **10**, 771–785 (2012).
51. Sasai, Y. Grow your own eye: biologists have coaxed cells to form a retina, a step toward growing replacement organs outside the body. *Sci. Am.* **307**, 44–49 (2012).
52. Ovando-Roche, P., Georgiadis, A., Smith, A. J., Pearson, R. A. & Ali, R. R. Harnessing the potential of human pluripotent stem cells and gene editing for the treatment of retinal degeneration. *Curr. Stem Cell Rep.* **3**, 112–123 (2017).
53. Kaewkhaw, R. et al. Transcriptome dynamics of developing photoreceptors in three-dimensional retina cultures recapitulates temporal sequence of human cone and rod differentiation revealing cell surface markers and gene networks. *Stem Cells* **33**, 3504–3518 (2015).

54. Busskamp, V. et al. Genetic reactivation of cone photoreceptors restores visual responses in retinitis pigmentosa. *Science* **329**, 413–417 (2010).
Demonstration of optogenetic vision restoration in human retinas.
55. Fradot, M. et al. Gene therapy in ophthalmology: validation on cultured retinal cells and explants from postmortem human eyes. *Hum. Gene Ther.* **22**, 587–593 (2011).
56. Dowling, J. E. *The Retina* (Harvard Univ. Press, Cambridge, 2012)
57. Izpisua Belmonte, J. C. et al. Brains, genes, and primates. *Neuron* **86**, 617–631 (2015).
58. Mitchell, J. F. & Leopold, D. A. The marmoset monkey as a model for visual neuroscience. *Neurosci. Res.* **93**, 20–46 (2015).
59. Sahel, J.-A. & Roska, B. Gene therapy for blindness. *Annu. Rev. Neurosci.* **36**, 467–488 (2013).
60. Carter, B. J. Adeno-associated virus and the development of adeno-associated virus vectors: a historical perspective. *Mol. Ther.* **10**, 981–989 (2004).
61. Ali, R. R. et al. Gene transfer into the mouse retina mediated by an adeno-associated viral vector. *Hum. Mol. Genet.* **5**, 591–594 (1996).
The first AAV gene transfer to the retina of mice.
62. Flannery, J. G. et al. Efficient photoreceptor-targeted gene expression in vivo by recombinant adeno-associated virus. *Proc. Natl Acad. Sci. USA* **94**, 6916–6921 (1997).
63. Planul, A. & Dalkara, D. Vectors and gene delivery to the retina. *Annu. Rev. Vis. Sci.* **3**, 121–140 (2017).
64. Madigan, V. J. & Asokan, A. Engineering AAV receptor footprints for gene therapy. *Curr. Opin. Virol.* **18**, 89–96 (2016).
65. Chaffiol, A. et al. A new promoter allows optogenetic vision restoration with enhanced sensitivity in macaque retina. *Mol. Ther.* **25**, 2546–2560 (2017).
66. Jacobson, S. G. et al. Gene therapy for Leber congenital amaurosis caused by RPE65 mutations: safety and efficacy in 15 children and adults followed up to 3 years. *Arch. Ophthalmol.* **130**, 9–24 (2012).
67. Wang, S. et al. Non-invasive, focused ultrasound-facilitated gene delivery for optogenetics. *Sci. Rep.* **7**, 39955 (2017).
68. Wan, C., Li, F. & Li, H. Gene therapy for ocular diseases meditated by ultrasound and microbubbles. *Mol. Med. Rep.* **12**, 4803–4814 (2015).
69. Touchard, E. et al. Non-viral gene therapy for GDNF production in RCS rat: the crucial role of the plasmid dose. *Gene Ther.* **19**, 886–898 (2012).
70. MacLaren, R. E. et al. Retinal repair by transplantation of photoreceptor precursors. *Nature* **444**, 203–207 (2006).
71. Pearson, R. A. et al. Restoration of vision after transplantation of photoreceptors. *Nature* **485**, 99–103 (2012).
72. MacLaren, R. E. Cone fusion confusion in photoreceptor transplantation. *Stem Cell Investig.* **4**, 71 (2017).
73. Singh, M. S. et al. Transplanted photoreceptor precursors transfer proteins to host photoreceptors by a mechanism of cytoplasmic fusion. *Nat. Commun.* **7**, 13537 (2016).
74. Pearson, R. A. et al. Donor and host photoreceptors engage in material transfer following transplantation of post-mitotic photoreceptor precursors. *Nat. Commun.* **7**, 13029 (2016).
75. Santos-Ferreira, T. et al. Retinal transplantation of photoreceptors results in donor–host cytoplasmic exchange. *Nat. Commun.* **7**, 13028 (2016).
76. Hertz, J. et al. Survival and integration of developing and progenitor-derived retinal ganglion cells following transplantation. *Cell Transplant.* **23**, 855–872 (2014).
77. Venugopalan, P. et al. Transplanted neurons integrate into adult retinas and respond to light. *Nat. Commun.* **7**, 10472 (2016).
78. Sahel, J. A. et al. Mitogenic effects of excitatory amino acids in the adult rat retina. *Exp. Eye Res.* **53**, 657–664 (1991).
79. Jorstad, N. L. et al. Stimulation of functional neuronal regeneration from Müller glia in adult mice. *Nature* **548**, 103–107 (2017).
80. Benowitz, L. & Yin, Y. Rewiring the injured CNS: lessons from the optic nerve. *Exp. Neurol.* **209**, 389–398 (2008).
81. Lim, J.-H. A. et al. Neural activity promotes long-distance, target-specific regeneration of adult retinal axons. *Nat. Neurosci.* **19**, 1073–1084 (2016).
82. Sun, F. et al. Sustained axon regeneration induced by co-deletion of PTEN and SOCS3. *Nature* **480**, 372–375 (2011).
83. Laha, B., Stafford, B. K. & Huberman, A. D. Regenerating optic pathways from the eye to the brain. *Science* **356**, 1031–1034 (2017).
84. Yue, L., Weiland, J. D., Roska, B. & Humayun, M. S. Retinal stimulation strategies to restore vision: fundamentals and systems. *Prog. Retin. Eye Res.* **53**, 21–47 (2016).
85. Scholl, H. P. N. et al. Emerging therapies for inherited retinal degeneration. *Sci. Transl. Med.* **8**, 368rv6 (2016).
86. Stingl, K. et al. Interim results of a multicenter trial with the new electronic subretinal implant Alpha AMS in 15 patients blind from inherited retinal degenerations. *Front. Neurosci.* **11**, 445 (2017).
87. Stronks, H. C. & Dagnelie, G. The functional performance of the Argus II retinal prosthesis. *Expert Rev. Med. Devices* **11**, 23–30 (2014).
88. Mullin, E. A company is reviving efforts to make a bionic eye brain implant for the blind. *MIT Technol. Rev*. https://www.technologyreview.com/s/608844/blind-patients-to-test-bionic-eye-brain-implants/ (2017).
89. Bi, A. et al. Ectopic expression of a microbial-type rhodopsin restores visual responses in mice with photoreceptor degeneration. *Neuron* **50**, 23–33 (2006).
90. Lagali, P. S. et al. Light-activated channels targeted to ON bipolar cells restore visual function in retinal degeneration. *Nat. Neurosci.* **11**, 667–675 (2008).
91. Polosukhina, A. et al. Photochemical restoration of visual responses in blind mice. *Neuron* **75**, 271–282 (2012).
92. Busskamp, V., Picaud, S., Sahel, J. A. & Roska, B. Optogenetic therapy for retinitis pigmentosa. *Gene Ther.* **19**, 169–175 (2012).
93. Berndt, A., Yizhar, O., Gunaydin, L. A., Hegemann, P. & Deisseroth, K. Bi-stable neural state switches. *Nat. Neurosci.* **12**, 229–234 (2009).
94. Hartong, D. T., Berson, E. L. & Dryja, T. P. Retinitis pigmentosa. *Lancet* **368**, 1795–1809 (2006).
95. den Hollander, A. I., Roepman, R., Koenekoop, R. K. & Cremers, F. P. M. Leber congenital amaurosis: genes, proteins and disease mechanisms. *Prog. Retin. Eye Res.* **27**, 391–419 (2008).
96. Dobelle, W. H., Mladejovsky, M. G. & Girvin, J. P. Artifical vision for the blind: electrical stimulation of visual cortex offers hope for a functional prosthesis. *Science* **183**, 440–444 (1974).
97. Bourne, D. et al. Whole-eye transplantation: a look into the past and vision for the future. *Eye (Lond.)* **31**, 179–184 (2017).
98. Miller, J. W. Beyond VEGF—the Weisenfeld lecture. *Invest. Ophthalmol. Vis. Sci.* **57**, 6911–6918 (2016).
99. Schwartz, S. D., Tan, G., Hosseini, H. & Nagiel, A. Subretinal transplantation of embryonic stem cell-derived retinal pigment epithelium for the treatment of macular degeneration: an assessment at 4 years. *Invest. Ophthalmol. Vis. Sci.* **57**, ORSFc1–ORSFc9 (2016).
100. Parker, M. A. et al. Test–retest variability of functional and structural parameters in patients with Stargardt disease participating in the SAR422459 gene therapy trial. *Transl. Vis. Sci. Technol.* **5**, 10 (2016).
101. Byrne, L. C. et al. Viral-mediated RdCVF and RdCVFL expression protects cone and rod photoreceptors in retinal degeneration. *J. Clin. Invest.* **125**, 105–116 (2015).
102. Léveillard, T. et al. Identification and characterization of rod-derived cone viability factor. *Nat. Genet.* **36**, 755–759 (2004).
Discovery of a factor for photoreceptor neuroprotection.
103. Mohand-Said, S. et al. Normal retina releases a diffusible factor stimulating cone survival in the retinal degeneration mouse. *Proc. Natl Acad. Sci. USA* **95**, 8357–8362 (1998).
104. Aït-Ali, N. et al. Rod-derived cone viability factor promotes cone survival by stimulating aerobic glycolysis. *Cell* **161**, 817–832 (2015).
105. Venkatesh, A. et al. Activated mTORC1 promotes long-term cone survival in retinitis pigmentosa mice. *J. Clin. Invest.* **125**, 1446–1458 (2015).
106. Punzo, C., Kornacker, K. & Cepko, C. L. Stimulation of the insulin/mTOR pathway delays cone death in a mouse model of retinitis pigmentosa. *Nat. Neurosci.* **12**, 44–52 (2009).
Discovery of a factor for photoreceptor neuroprotection.
107. Xiong, W., MacColl Garfinkel, A. E., Li, Y., Benowitz, L. I. & Cepko, C. L. NRF2 promotes neuronal survival in neurodegeneration and acute nerve damage. *J. Clin. Invest.* **125**, 1433–1445 (2015).
108. Cepko, C. & Punzo, C. Cell metabolism: sugar for sight. *Nature* **522**, 428–429 (2015).
109. Krol, J. & Roska, B. Rods feed cones to keep them alive. *Cell* **161**, 706–708 (2015).
110. Léveillard, T. & Sahel, J.-A. Metabolic and redox signaling in the retina. *Cell. Mol. Life Sci.* **74**, 3649–3665 (2016).
111. Campochiaro, P. A. et al. Is there excess oxidative stress and damage in eyes of patients with retinitis pigmentosa? *Antioxid. Redox Signal.* **23**, 643–648 (2015).
112. Kayser, S. et al. Reduced central retinal artery blood flow is related to impaired central visual function in retinitis pigmentosa patients. *Curr. Eye Res.* **42**, 1503–1510 (2017).
113. Chaikin, L., Kashiwa, K., Bennet, M., Papastergiou, G. & Gregory, W. Microcurrent stimulation in the treatment of dry and wet macular degeneration. *Clin. Ophthalmol.* **9**, 2345–2353 (2015).
114. Schatz, A. et al. Transcorneal electrical stimulation for patients with retinitis pigmentosa: a prospective, randomized, sham-controlled follow-up study over 1 year. *Invest. Ophthalmol. Vis. Sci.* **58**, 257–269 (2017).
115. Age-Related Eye Disease Study Research Group. A randomized, placebo-controlled, clinical trial of high-dose supplementation with vitamins C and E, beta carotene, and zinc for age-related macular degeneration and vision loss: AREDS report no. *8. Arch. Ophthalmol.* **119**, 1417–1436 (2001).
116. Jurkute, N. & Yu-Wai-Man, P. Leber hereditary optic neuropathy: bridging the translational gap. *Curr. Opin. Ophthalmol.* **28**, 403–409 (2017).
117. Smirnakis, S. M. Probing human visual deficits with functional magnetic resonance imaging. *Annu. Rev. Vis. Sci.* **2**, 171–195 (2016).
118. Bainbridge, J. W. B. et al. Long-term effect of gene therapy on Leber's congenital amaurosis. *N. Engl. J. Med.* **372**, 1887–1897 (2015).
119. Jacobson, S. G. et al. Improvement and decline in vision with gene therapy in childhood blindness. *N. Engl. J. Med.* **372**, 1920–1926 (2015).
120. Rizzo, J. F. III & Ayton, L. N. Psychophysical testing of visual prosthetic devices: a call to establish a multi-national joint task force. *J. Neural Eng.* **11**, 020301 (2014).
121. Finger, R. P. et al. Developing the impact of Vision Impairment-Very Low Vision (IVI-VLV) questionnaire as part of the LoVADA protocol. *Invest. Ophthalmol. Vis. Sci.* **55**, 6150–6158 (2014).
122. Jeter, P. E., Rozanski, C., Massof, R., Adeyemo, O. & Dagnelie, G. Development of the Ultra-Low Vision Visual Functioning Questionnaire (ULV-VFQ). *Transl. Vis. Sci. Technol.* **6**, 11 (2017).
123. Authié, C. N., Berthoz, A., Sahel, J.-A. & Safran, A. B. Adaptive gaze strategies for locomotion with constricted visual field. *Front. Hum. Neurosci.* **11**, 387 (2017).
124. Cattaneo, Z. & Vecchi, T. *Blind Vision* (MIT Press, Cambridge, 2011).
125. Irvine, D. et al. Tablet and smartphone accessibility features in the low vision rehabilitation. *Neuroophthalmology* **38**, 53–59 (2014).

126. Robinson, J. L., Braimah Avery, V., Chun, R., Pusateri, G. & Jay, W. M. Usage of accessibility options for the iPhone and iPad in a visually impaired population. *Semin. Ophthalmol.* **32**, 163–171 (2017).
127. Alter, C. Facebook is developing technology that can describe pictures to blind people. *Time* http://time.com/4099204/facebook-artificial-intelligence-blind-pictures/ (2015).
128. Striem-Amit, E., Guendelman, M. & Amedi, A. 'Visual' acuity of the congenitally blind using visual-to-auditory sensory substitution. *PLoS One* **7**, e33136 (2012).
129. Striem-Amit, E., Cohen, L., Dehaene, S. & Amedi, A. Reading with sounds: sensory substitution selectively activates the visual word form area in the blind. *Neuron* **76**, 640–652 (2012).
130. Thaler, L. & Goodale, M. A. Echolocation in humans: an overview. *Wiley Interdiscip. Rev. Cogn. Sci.* **7**, 382–393 (2016).
131. Köberlein, J., Beifus, K., Schaffert, C. & Finger, R. P. The economic burden of visual impairment and blindness: a systematic review. *BMJ Open* **3**, e003471 (2013).
132. Romanov, R. A. et al. Molecular interrogation of hypothalamic organization reveals distinct dopamine neuronal subtypes. *Nat. Neurosci.* **20**, 176–188 (2017).

Acknowledgements We thank S. Trenholm for discussions, comments and advice on the manuscript; A. Drinnenberg, D. Dalkara, S. Picaud, I. Audo, K. Marazova, S. Oakeley and P. King for comments on the manuscript; and P. Maloca for images for Fig. 1, A. Drinnenberg for drawings for Fig. 2, and V. Juvin (http://www.sciartwork.com) for images for Fig. 3.

Reviewer information *Nature* thanks G. Dagnelie, J. Demb and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions B.R. and J.A.-S. discussed and wrote the manuscript.

Competing interests J.A.-S. has financial interests in GenSight Biologics, Chronocam, Chronolife, Pixium Vision, Tilak Healthcare, and Sparing Vision.

Additional information
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to B.R. and J.A.-S.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

ARTICLE

https://doi.org/10.1038/s41586-018-0094-2

137 ancient human genomes from across the Eurasian steppes

Peter de Barros Damgaard[1], Nina Marchi[2], Simon Rasmussen[3], Michaël Peyrot[4], Gabriel Renaud[1], Thorfinn Korneliussen[1,5], J. Víctor Moreno-Mayar[1], Mikkel Winther Pedersen[5], Amy Goldberg[6], Emma Usmanova[7], Nurbol Baimukhanov[8], Valeriy Loman[7], Lotte Hedeager[9], Anders Gorm Pedersen[3], Kasper Nielsen[3,50], Gennady Afanasiev[10], Kunbolot Akmatov[11], Almaz Aldashev[12], Ashyk Alpaslan[11], Gabit Baimbetov[8], Vladimir I. Bazaliiskii[13], Arman Beisenov[14], Bazartseren Boldbaatar[15], Bazartseren Boldgiv[16], Choduraa Dorzhu[17], Sturla Ellingvag[18], Diimaajav Erdenebaatar[19], Rana Dajani[20,21], Evgeniy Dmitriev[7], Valeriy Evdokimov[7], Karin M. Frei[22], Andrey Gromov[23], Alexander Goryachev[24], Hakon Hakonarson[25], Tatyana Hegay[26], Zaruhi Khachatryan[27], Ruslan Khaskhanov[28], Egor Kitov[14,29], Alina Kolbina[30], Tabaldiev Kubatbek[11], Alexey Kukushkin[7], Igor Kukushkin[7], Nina Lau[31], Ashot Margaryan[1,32], Inga Merkyte[33], Ilya V. Mertz[34], Viktor K. Mertz[34], Enkhbayar Mijiddorj[19], Vyacheslav Moiyesev[23], Gulmira Mukhtarova[35], Bekmukhanbet Nurmukhanbetov[35], Z. Orozbekova[36], Irina Panyushkina[37], Karol Pieta[38], Václav Smrčka[39], Irina Shevnina[40], Andrey Logvin[40], Karl-Göran Sjögren[41], Tereza Štolcová[38], Kadicha Tashbaeva[42], Alexander Tkachev[43], Turaly Tulegenov[35], Dmitriy Voyakin[24], Levon Yepiskoposyan[27], Sainbileg Undrakhbold[16], Victor Varfolomeev[7], Andrzej Weber[44], Nikolay Kradin[45,46], Morten E. Allentoft[1], Ludovic Orlando[1,47], Rasmus Nielsen[1,48], Martin Sikora[1], Evelyne Heyer[2], Kristian Kristiansen[41] & Eske Willerslev[1,5,49]*

For thousands of years the Eurasian steppes have been a centre of human migrations and cultural change. Here we sequence the genomes of 137 ancient humans (about 1× average coverage), covering a period of 4,000 years, to understand the population history of the Eurasian steppes after the Bronze Age migrations. We find that the genetics of the Scythian groups that dominated the Eurasian steppes throughout the Iron Age were highly structured, with diverse origins comprising Late Bronze Age herders, European farmers and southern Siberian hunter-gatherers. Later, Scythians admixed with the eastern steppe nomads who formed the Xiongnu confederations, and moved westward in about the second or third century BC, forming the Hun traditions in the fourth–fifth century AD, and carrying with them plague that was basal to the Justinian plague. These nomads were further admixed with East Asian groups during several short-term khanates in the Medieval period. These historical events transformed the Eurasian steppes from being inhabited by Indo-European speakers of largely West Eurasian ancestry to the mostly Turkic-speaking groups of the present day, who are primarily of East Asian ancestry.

The Eurasian steppes stretch about 8,000 km from Hungary and Romania in the west to Mongolia and northeastern China in the east. These regions have, in the past four to five millennia, been dominated first by Iranian- and later by Turkic- and Mongolic-speaking nomadic groups with herding and warrior economies. To understand the population genetic processes associated with the linguistic and cultural changes of the steppes after the Bronze Age migrations[1–3], we sequenced 137 ancient genomes—to about 1× average depth (see Supplementary Tables 1, 2)—from Europe to Mongolia and the Altai to Tian Shan mountains; these genomes covered approximately 4,000 years (about 2500 BC–AD 1500) (Fig. 1). A list of the population labels used throughout this Article can be found in Supplementary Table 3.

---

[1]Center for GeoGenetics, Natural History Museum of Denmark, University of Copenhagen, Copenhagen, Denmark. [2]Eco-anthropologie et Ethnobiologie, Muséum national d'Histoire naturelle, CNRS, Université Paris Diderot, Paris, France. [3]Department of Bio and Health Informatics, Technical University of Denmark, Lyngby, Denmark. [4]Leiden University Centre for Linguistics, Leiden University, Leiden, The Netherlands. [5]Department of Zoology, University of Cambridge, Cambridge, UK. [6]Department of Biology, Stanford University, Stanford, CA, USA. [7]Buketov Karaganda State University, Saryarka Archaeological Institute, Karaganda, Kazakhstan. [8]Shejire DNA, Almaty, Kazakhstan. [9]Department of Archaeology, Conservation and History, University of Oslo, Oslo, Norway. [10]Department of Theory and Methods, Institute of Archaeology Russian Academy of Sciences, Moscow, Russia. [11]Department of History, Kyrgyzstan-Turkey Manas University, Bishkek, Kyrgyzstan. [12]National Academy of Sciences of Kyrgyzstan, Bishkek, Kyrgyzstan. [13]Department of History, Irkutsk State University, Irkutsk, Russia. [14]A. Kh. Margulan Institute of Archaeology, Almaty, Kazakhstan. [15]Laboratory of Virology, Institute of Veterinary Medicine, Mongolian University of Life Sciences, Ulaanbaatar, Mongolia. [16]Department of Biology, School of Arts and Sciences, National University of Mongolia, Ulaanbaatar, Mongolia. [17]Departament of Biology and Ecology, Tuvan State University, Kyzyl, Russia. [18]The Explico Foundation, Floro, Norway. [19]Department of Archaeology, Ulaanbaatar State University, Ulaanbaatar, Mongolia. [20]Department of Biology and Biotechnology, Hashemite University, Zarqa, Jordan. [21]Radcliffe Institute for Advanced Study, Harvard University, Cambridge, MA, USA. [22]Unit for Environmental Archaeology and Materials Science, National Museum of Denmark, Copenhagen, Denmark. [23]Peter the Great Museum of Anthropology and Ethnography (Kunstkamera) RAS, St. Petersburg, Russia. [24]Archaeological Expertise LLC, Almaty, Kazakhstan. [25]Center for Applied Genomics, The Children's Hospital of Philadelphia, Philadelphia, PA, USA. [26]Republican Scientific Center of Immunology, Ministry of Public Health, Tashkent, Uzbekistan. [27]Department of Bioengineering, Bioinformatics and Molecular Biology, Russian-Armenian University, Yerevan, Armenia. [28]Complex Research Institute of the Russian Academy of Sciences, Grozny, Russia. [29]Institute of Ethnology and Anthropology, Russian Academy of Science, Moscow, Russia. [30]Kostanay Regional Local History Museum, Kostanay, Kazakhstan. [31]Centre for Baltic and Scandinavian Archaeology, Schleswig, Germany. [32]Laboratory of Ethnogenomics, Institute of Molecular Biology, National Academy of Sciences of Armenia, Yerevan, Armenia. [33]Saxo-Institute, University of Copenhagen, Copenhagen, Denmark. [34]Center for Archaeological Research, S. Toraighyrov Pavlodar State University, Pavlodar, Kazakhstan. [35]The State Historical and Cultural Reserve-Museum (ISSYK), Almaty, Kazakhstan. [36]Institute of Archeology and Ethnography of the Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia. [37]University of Arizona, Laboratory of Tree-Ring Research, Tucson, AZ, USA. [38]Institute of Archaeology of the Slovak Academy of Sciences, Nitra, Slovakia. [39]Institute for History of Medicine and Foreign Languages, First Faculty of Medicine, Charles University, Prague, Czech Republic. [40]Archaeological Laboratory, Kostanay State University, Kostanay, Kazakhstan. [41]Department of Historical Studies, University of Gothenburg, Gothenburg, Sweden. [42]Institute of History and Cultural Heritage of National Academy of Sciences, Bishkek, Kyrgyzstan. [43]Institute of Problems Development of the North Siberian Branch of the Russian Academy of Sciences, Tyumen, Russia. [44]Department of Anthropology, University of Alberta, Edmonton, Alberta, Canada. [45]Institute of History, Archaeology and Ethnology, Far-Eastern Branch of the Russian Academy of Sciences, Ulan-Ude, Russia. [46]Institute of Mongolian, Buddhist, and Tibetan Studies, Siberian Branch of the Russian Academy of Sciences, Ulan-Ude, Russia. [47]Laboratoire d'Anthropobiologie Moléculaire et d'Imagerie de Synthèse, Université de Toulouse, Université Paul Sabatier, Toulouse, France. [48]Departments of Integrative Biology and Statistics, University of Berkeley, Berkeley, CA, USA. [49]Wellcome Trust Sanger Institute, Hinxton, UK. [50]Present address: Carlsberg Research Laboratory, Copenhagen, Denmark. *e-mail: ewillerslev@snm.ku.dk

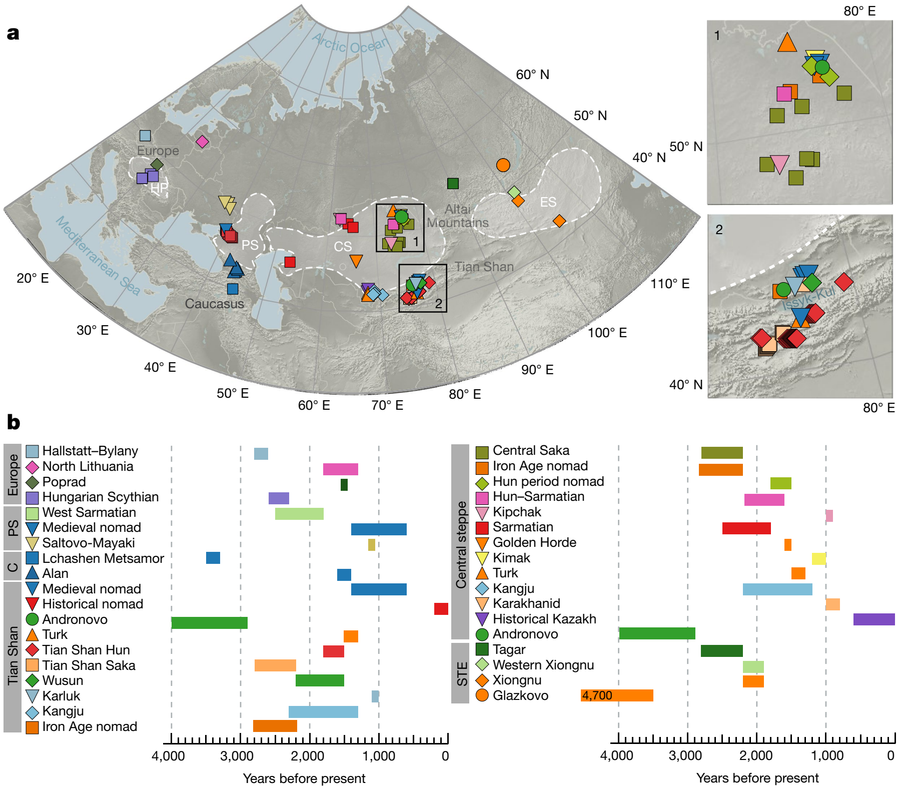


Fig. 1 | Cultural and geographical presentation of the ancient samples. **a**, Geographical distribution of samples. Symbols correspond to samples of a specific age: circle, Bronze Age; square, Iron Age; diamond, Hun period; triangle upwards, Turk period; triangle downwards, Medieval period. **b**, Each symbol has been sorted according to geographical region highlighted on the map in **a**, and given in the grey boxes in **b**. C, Caucasus; CS, central steppe; ES, eastern steppe; HP, Hungarian plains; PS, Pontic steppe; STE, Siberia, Tungus and eastern steppe (STE is not marked on the map in **a**, but includes steppe and non-steppe areas).

Furthermore, we genotyped 502 individuals of 16 self-reported ethnicities from across Central Asia, Altai, Siberia and the Caucasus (Supplementary Table 4 and Supplementary Information section 5). In the process, we tested differential ancient DNA preservation in organic contrasted mineral substrate (Supplementary Information section 6), and generated 83 new accelerator mass spectrometry dates (Supplementary Information section 11).

The genomic origins of the Scythian confederations

Between about 800 and 200 BC, the Eurasian steppes became dominated by the Iranian-speaking Scythians. This confederation was divided into geographically distinct groups, but was united by similarities in cultural expression[4]. However, the origins and population structure of the Scythians remain contested, as can be summarized in three competing models: (1) the Scythians deriving from a single source originating in the northern Caucasus or steppe region[5–7]; (2) an origin in southern Siberia or east-central Asia, moving westwards[8,9]; and finally (3) the Scythians being a product of multiple transitions taking place locally, involving social and cultural borrowing in combination with gradual, small-scale human movements[10–13].

Using principal component analysis (PCA) and ADMIXTURE analyses (Fig. 2 and Extended Data Fig. 1), we observe a clear separation between two groups of Iron Age Scythians: the Hungarian Scythians and the Inner Asian Sakas. Furthermore, we find fine-scaled structure within the Inner Asian Sakas that separates (1) the populations associated with the 'Tagar' culture of southern Siberia, (2) the 'Central Sakas' of the central steppe—most of whom have been described as belonging to the Tasmola culture (Supplementary Information section 3)—and (3) the 'Tian Shan Sakas' of the Tian Shan mountain range (see map in Fig. 1). These differences reflect the confederal nature of the Scythian organization.

Recent genetic models suggested the presence of Yamnaya and/or Afanasievo ancestry in Scythians[11], which we assessed here using a new set of outgroups that enabled us to distinguish between Early and Late Bronze Age steppe ancestry (Supplementary Information section 3.6). We find that the Late Bronze Age herders are a better genetic source for the West Eurasian ancestry in Scythians than are Early Bronze Age Yamnaya or Afanasievo, the key difference being their European farmer ancestry (Supplementary Table 5). Using ADMIXTURE models[14] we also illustrate the shared ancestry between Neolithic farmers (from Anatolia or Europe), Late Bronze Age herders and Iron Age steppe nomads that is not shared with Yamnaya herders (Extended Data Fig. 2 and Supplementary Fig. 163). These findings are consistent with archaeological models.

Using *D*-statistics (Supplementary Information section 3.7), we then characterized the sources of admixture into the various Scythian groups relative to the Late Bronze Age steppe herders. We find that Hungarian Scythians had relatively increased European farmer ancestry (Extended Data Fig. 3) and show no signs of gene flow from Inner Asian groups. Conversely, Inner Asian Sakas show relatively increased southern Siberian hunter-gatherer ancestry with the strongest gene flow observed into the Central Sakas. This East Asian admixture is also reflected in the negative admixture f_3 values, indicating that Late Bronze Age pastoralists and southern Siberian hunter-gatherers are excellent proxies for the admixing populations (Extended Data Fig. 4). We confirm the differences between these Iron Age steppe groups through *D*-statistics (Supplementary Information section 3.7). The increase in Neolithic Iranian ancestry in the Tian Shan Sakas is significant when compared to Central Sakas; the Tagar display increased eastern hunter-gatherer (EHG) ancestry compared to all other Scythians. Lastly, the high genetic differentiation between western and eastern Scythians is emphasized by observing higher fixation index (F_{ST}) values between Hungarian Scythians and all Inner Asian Sakas (F_{ST} ranges from 0.24 to 0.3) than observed among the different Inner Asian Sakas groups (F_{ST} ranges from 0.15 to 0.2) (Supplementary Table 6).

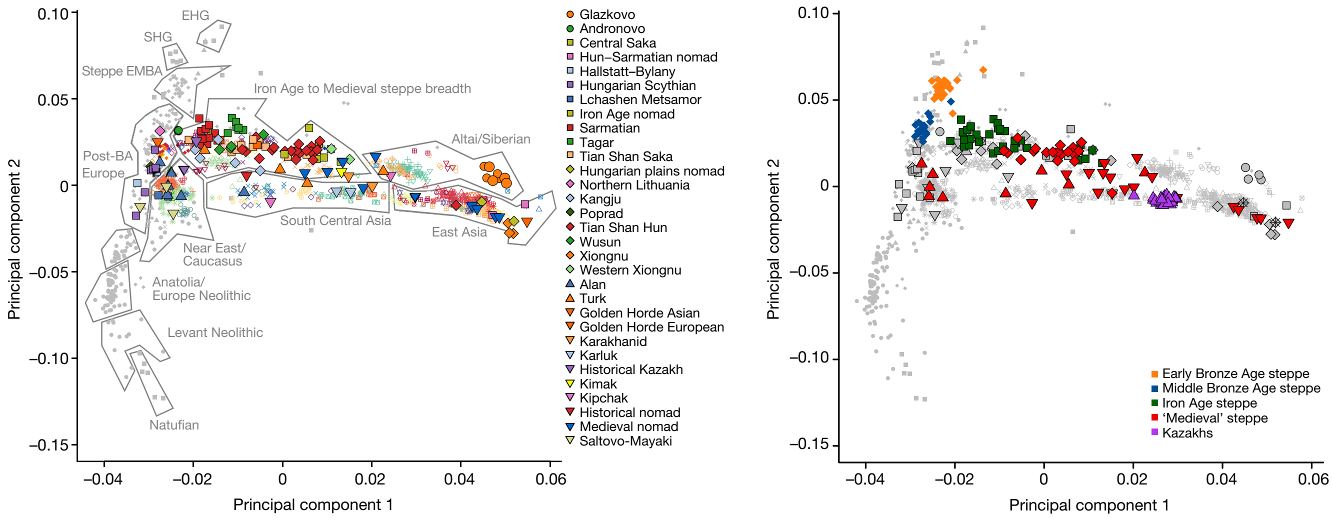


Fig. 2 | Principal component analyses. The principal components 1 and 2 were plotted for the ancient data analysed with the present-day data (no projection bias) using 502 individuals at 242,406 autosomal SNP positions. Dimension 1 explains 3% of the variance and represents a gradient stretching from Europe to East Asia. Dimension 2 explains 0.6% of the variance, and is a gradient mainly represented by ancient DNA starting from a 'basal-rich' cluster of Natufian hunter-gatherers and ending with EHGs. BA, Bronze Age; EMBA, Early-to-Middle Bronze Age; SHG, Scandinavian hunter-gatherers.

The qpAdm modelling[1] of this ancient genomic dataset is consistent with these findings. The Central Sakas can be modelled as a simple two-way mixture of Late Bronze Age pastoralists and southern Siberian hunter-gatherers, with almost equal proportions of Bronze Age herder (56%) and southern Siberian hunter-gatherer ancestry (44%). The southern Siberian Tagar show unequal ancestry contributions from Bronze Age herders (83.5%) and southern Siberian hunter-gatherers (7.5%), as well as an additional contribution of Mal'ta (MA1 individual)-like ancestry (9%), indicating differences in the sources of hunter-gatherer admixture across the Sakas. The Saka population of the Tian Shan mountains displays a high proportion of Late Bronze Age steppe herder ancestry (70%) followed by southern Siberian hunter-gatherer ancestry (25%), and also an additional 5% ancestry coming from a source related to a Neolithic population from Iran. Taken together, our data do not support the recent mtDNA-based claim of extensive gene flow between the different Scythian groups[11], but instead indicate admixture between populations of Late Bronze Age herder descent and various local groups, consistent with the multiple origins model (model 3 described above).

Our data show that the culturally similar Scythians represented genetically structured groups within the Eurasian steppes. In particular, the Siberian Tagar, Central Sakas and the Tian Shan Sakas were Scythian groups that arose through admixture between Late Bronze Age pastoral groups and Inner Asian hunter-gatherers, in contrast to the Hungarian Scythians who received gene flow from farming groups within Europe. The additional gene flow from a source related to the Neolithic Iranians detected in the Tian Shan Sakas suggests that southern steppe nomads also interacted with the civilization of the Bactria–Margiana archaeological complex of present-day eastern Turkmenistan.

The Xiongnu and the Hunnic expansions

Turkic language elements arguably first emerged among the Xiongnu nomads[15], a confederation of several nomadic tribes who occupied the eastern steppe from the third century BC. They are believed to be of East Asian ancestry[16,17], although ancient Y-chromosomal data have indicated a possibly heterogeneous population admixed with central steppe nomads[18]. Huns (third–fifth century AD) have previously been argued to derive directly from the Xiongnu[19], although others have claimed that there is no evidence connecting the two groups[20]. It is commonly believed that the Huns spread westward, disseminating Turkic languages throughout Central Asia at the cost of Iranian languages. It is known that the expansion of the Xiongnu nomads affected the movements of other cultural groups from the south-eastern side of the Tian Shan Mountains, such as the Wusun and Kangju, whose genetic ancestries have so far remained unknown. It has tentatively been suggested on the basis of the archaeological record that they belonged to the Iranian-speaking branch of the Indo-European language family[21].

Principal component analyses and *D*-statistics suggest that the Xiongnu individuals belong to two distinct groups, one being of East Asian origin and the other presenting considerable admixture levels with West Eurasian sources (Fig. 2 and Extended Data Figs. 1, 5; in Fig. 2 these are labelled 'Xiongnu' and 'western Xiongnu', respectively). We find that Central Sakas are accepted as a source for these 'western-admixed' Xiongnu in a single-wave model. Consistent with this finding, no East Asian gene flow is detected compared to Central Sakas as these form a clade with respect to the East Asian Xiongnu in a *D*-statistic, and cluster closely together in the PCA (Fig. 2).

We used *D*-statistics (Supplementary Information section 3.7) to investigate the genetic relationship between Iron Age nomads, the East Asian Xiongnu and the early Huns of the Tian Shan. We find that the Huns have increased shared drift with West Eurasians compared to the Xiongnu (Extended Data Fig. 6). We tested for patterns of shared drift between the Xiongnu and the Wusun, the preceding Sakas and the slightly later Huns (second century AD). We find that both the earlier Sakas and the later Huns have more East Asian ancestry than the Wusun. This is also apparent from model-based clustering and PCA (Extended Data Fig. 7). Similar results are seen with the contemporaneous and later Kangju groups that—as did the Wusun—re-emerged into the central steppe from south-east of the Tian Shan mountains. In addition, both groups require a Neolithic Iranian-related source for modelling ancestral proportions in the qpAdm framework (Supplementary Table 7), together with Late Bronze Age pastoralists and the southern Siberian hunter-gatherers. We therefore suspect that the Wusun and Kangju groups are descendants of Bronze Age pastoralists that interacted with the civilization of the Bactria–Margiana archaeological complex in southern Uzbekistan and eastern Turkmenistan, yet remained much less admixed with East Asians than did the Iron Age steppe Sakas.

Overall, our data show that the Xiongnu confederation was genetically heterogeneous, and that the Huns emerged following minor male-driven East Asian gene flow into the preceding Sakas that they invaded (see Supplementary Information section 3.6 for sex-biased admixture rates). As such our results support the contention that the

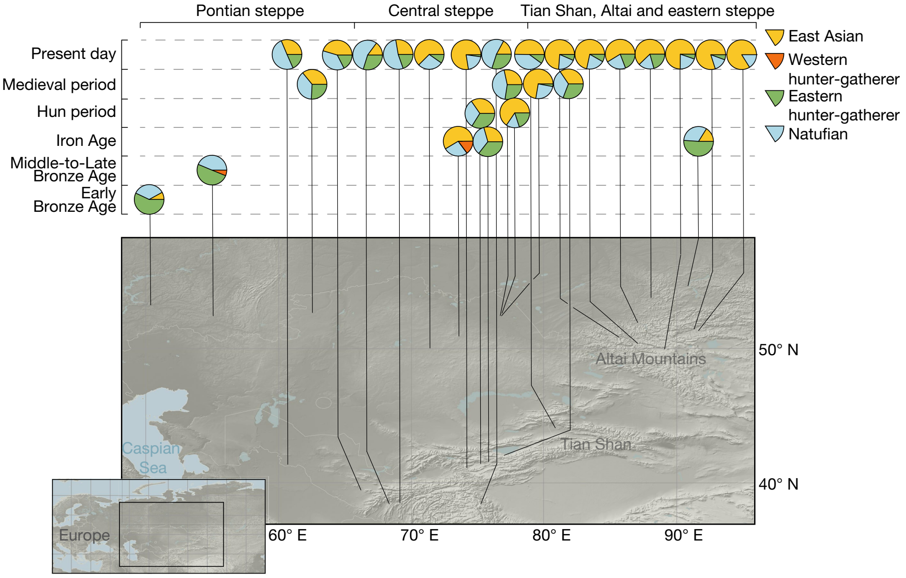


Fig. 3 | QpAdm results depicting the changes in ancestry across time in Central Asia. The changes reflect a gradual increase in East Asian ancestry in the central steppe nomads coupled to a decrease in ancestry associated with EHGs, starting at a high level in Yamnaya and finishing at a low level in present-day Kazakh and Kyrgyz individuals. The set of outgroups used is: Mbuti, Ust'Ishim, Clovis, Kostenki14 and Scandinavian hunter-gatherers.

disappearance of the Inner Asian Scythians and Sakas around two thousand years ago was a cultural transition that coincided with the westward migration of the Xiongnu. This Xiongnu invasion also led to the displacement of isolated remnant groups—related to Late Bronze Age pastoralists—that had remained on the south-eastern side of the Tian Shan mountains.

Repeated conquests and waves of East Asian impact

In the sixth century AD, the Hunnic Empire had been broken up and dispersed as the Turkic Khaganate assumed the military and political domination of the steppes[22,23]. Khaganates were steppe nomad political organizations that varied in size and became dominant during this period; they can be contrasted to the previous stateless organizations of the Iron Age[24]. The Turkic Khaganate was eventually replaced by a number of short-lived steppe cultures[25]. These included the Kipchak and the Tungusic Kimak populations, which spread southwards towards the Tian Shan mountains and westward towards the Ural mountains to form the Kimak Khaganate of the central steppe during the eighth to eleventh centuries AD[26]. During the eleventh century, the Kimak Khaganate was overthrown by local Kipchak groups, who in turn allied themselves with the Cuman of West Eurasia. Eventually the short-lived khaganates were overtaken by the Mongol Empire, which emerged through the unification of East Mongolian and Transbaikalian tribes and which expanded considerably during the rule of Genghis Khan in the thirteenth century AD[26,27] (Supplementary Information section 1).

We find evidence that elite soldiers associated with the Turkic Khaganate are genetically closer to East Asians than are the preceding Huns of the Tian Shan mountains (Supplementary Information section 3.7). We also find that one Turkic Khaganate-period nomad was a genetic outlier with pronounced European ancestries, indicating the presence of ongoing contact with Europe. Only one sample here represents Kimak nomads, and it does not show elevated East Asian ancestry (Supplementary Information section 3.7). During the Kipchak period in the eleventh century AD, the domination of the central steppe was allegedly assumed by another group originating from the geographical area of Tuva. We present genomic data from two individuals from this period, one of whom shows increased East Asian ancestry, whereas the other has pronounced European ancestry (samples DA23 and DA179, respectively, in Supplementary Information section 4). These individuals date to the Cuman–Kipchak alliance, which incorporated both the western and eastern steppe. For the period in which the region became incorporated into the Karakhanid Khanate—which encompassed present-day regions of Uzbekistan, Tajikistan, Kazakhstan and Kyrgyzstan—*D*-statistics identify a small influx of East Asian ancestry compared to the earlier Turk period. Consistent with this, nomads in the Karakhanid period are shifted towards East Asians compared to earlier Turks in the PCA plot (Fig. 2 and Extended Data Fig. 8). Additionally, we analysed ten culturally unaffiliated Medieval-period nomads, most of whom showed pronounced East Asian ancestry, albeit in very different proportions (Extended Data Fig. 8). We also find the presence of an individual of West Eurasian descent buried together with members of Jochi Khan's Golden Horde army from the Ulytau mountains (see Supplementary Information section 4: DA28 is East Asian and DA29 is European). This could suggest assimilation of distinct groups into the Medieval Golden Horde, but this individual may also represent a slave or a servant of West Eurasian descent attached to the service of the Golden Horde members.

These results suggest that Turkic cultural customs were imposed by an East Asian minority elite onto central steppe nomad populations, resulting in a small detectable increase in East Asian ancestry. However, we also find that steppe nomad ancestry in this period was extremely heterogeneous, with several individuals being genetically distributed at the extremes of the first principal component (Fig. 2) separating Eastern and Western descent. On the basis of this notable heterogeneity, we suggest that during the Medieval period steppe populations were exposed to gradual admixture from the east, while interacting with incoming West Eurasians. The strong variation is a direct window into ongoing admixture processes and the multi-ethnic cultural organization of this period.

Origins and spread of the Justinian plague

A few decades after the period of Hunnic-driven mobility across the Eurasian steppes, large areas of Europe were depopulated owing to the Justinian plague pandemic[28]. Although the first reports of the pandemic point to an outbreak in Egypt from where it is thought to have spread into Europe[29], the primordial origins of the Justinian plague remain unknown. The most basal strains of present-day plague (0.PE7 clade) have been found in Qinghai, south-east of the Tian Shan mountains[30], and the clade basal to the Justinian plague (0.ANT1) was found in Xinjiang in China, thus pointing to a possible Inner Asian origin of the Justinian plague.

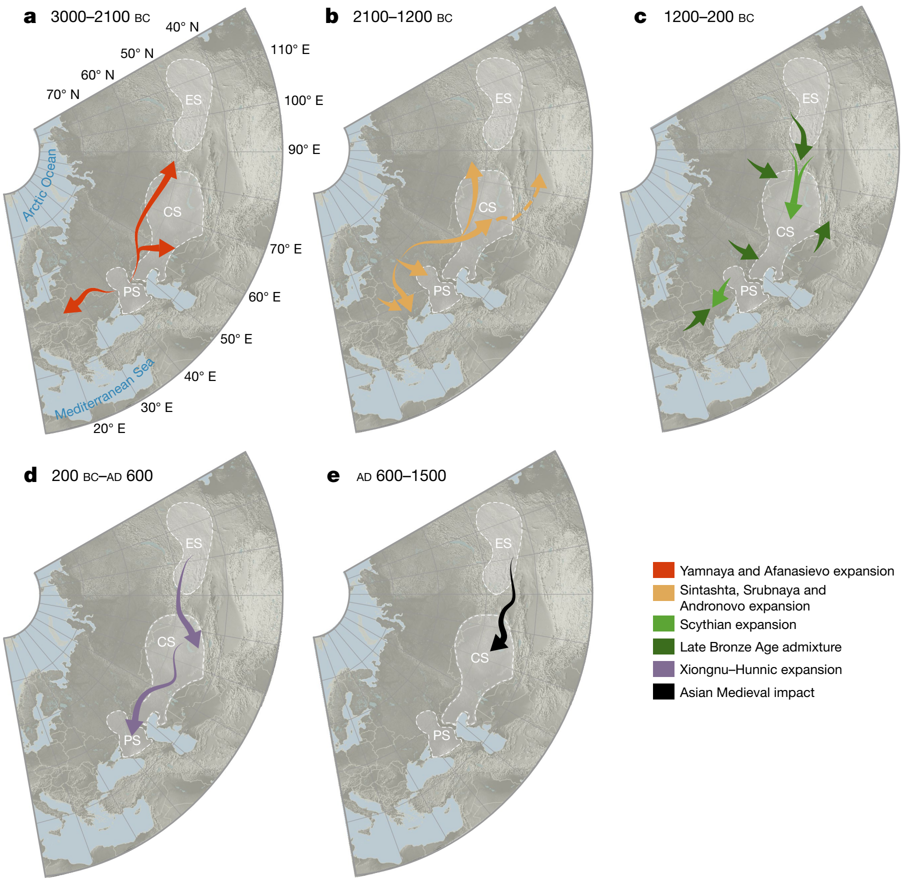


Fig. 4 | Summary map. Depictions of the five main migratory events associated with the genomic history of the steppe pastoralists from 3000 BC to the present. **a**, Depiction of Early Bronze Age migrations related to the expansion of Yamnaya and Afanasievo culture. **b**, Depiction of Late Bronze Age migrations related to the Sintashta and Andronovo horizons. **c**, Depiction of Iron Age migrations and sources of admixture. **d**, Depiction of Hun-period migrations and sources of admixture. **e**, Depiction of Medieval migrations across the steppes.

We find that two individuals, DA101 and DA147 (see Supplementary Information section 7), show detectable levels of *Yersinia pestis* DNA, compatible with the characterization of the full genome sequence at 8.7× and 0.24× coverage. The first individual (DA101) is a Hun from the Tian Shan mountains and dates to approximately AD 180, and the second individual (DA147) is from the Alan culture from North Ossetia and is estimated archaeologically to date to the sixth–ninth century AD. The genome of the *Y. pestis* strain DA101, which we name 0.ANT5, branches off from the main plague lineage just basal to the Justinian plague strain 0.ANT4, identified from an individual in Aschheim (Germany) and dated to about AD 530[29] (Extended Data Fig. 9). As expected, the Tian Shan strain contained the *ymt* gene reported to be missing in the more-ancestral Bronze Age plague strains[31]. The strain also displayed the loss of function mutations in *pde2*, *pde3*, *rcsA* and *ureD* that are required for flea transmission in the traditional ‘blocked flea’ model[32] (Extended Data Fig. 9). This, coupled with a fully functional plasminogen activator gene, indicates that the ‘Hunnic’ plague strain had full bubonic capability and flea transmissibility.

The fact that we find a higher number of strain-specific variants in the Aschheim strain is consistent with the difference in sampling time (approximately AD 180 versus approximately AD 530) and the potentially multiple replication cycles associated with pandemics[30]. This is supported by the substitution rate on the branch leading to the Aschheim strain being higher. Mutation rates in pathogens have been hypothesized to be affected by epidemics, not only because of natural selection but also owing to an increase in replication rate[30]. Therefore, our observation of an accelerated mutation rate is consistent with this hypothesis and supports the idea that the Ascheim strain was responsible for a major outbreak—the Justinian plague.

Given that the most basal strains of present-day plague (0.PE7 clade) originate in Qinghai[30] and the clade basal to the Justinian plague (0.ANT1) is from Xinjiang (China), two areas close to the Tian Shan mountains, we find provisional support for the hypothesis that the pandemic was brought to Europe towards the end of the Hunnic period through the Silk Road along the southern fringes of the steppes.

Discussion

The overall population history that formed the genetic composition of present-day steppe populations is illustrated in Fig. 3, in which we model the entire known ancient and present-day diversity of Inner Asia using the key ancestral groups. We also identify sex-specific admixture proportions in the Iron Age (Extended Data Fig. 10 and Supplementary Information section 3.6). In Fig. 4, we present the main migratory patterns. Our findings fit well with current insights from the historical linguistics of this region (Supplementary Information section 2). The steppes were probably largely Iranian-speaking in the first and second millennia BC. This is supported by the split of the Indo-Iranian linguistic branch into Iranian and Indian[33], the distribution of the Iranian languages, and the preservation of Old Iranian loanwords in Tocharian[34]. The wide distribution of the Turkic languages from Northwest China, Mongolia and Siberia in the east to Turkey and Bulgaria in the west implies large-scale migrations out of the homeland in Mongolia since about 2,000 years ago[35]. The diversification within the Turkic languages suggests that several waves of migration occurred[36] and, on the basis of the effect of local languages, gradual assimilation to local populations had previously been assumed[37]. The East Asian migration starting with the Xiongnu accords well with the hypothesis that early Turkic was the major language of Xiongnu groups[38]. Further migrations of East Asians westwards find a good linguistic correlate in the influence of Mongolian on Turkic and Iranian in the last millennium[39]. As such, the genomic history of the Eurasian steppes is the story of a gradual transition from Bronze Age pastoralists of West Eurasian ancestry towards mounted warriors of increased East Asian ancestry—a process that continued well into historical times.

Data availability

Sequence data were deposited in the European Nucleotide Archive (ENA) under accession number PRJEB20658 (ERP022829). Single nucleotide polymorphism data for present-day populations are available, after ethical validation, from the European Genome-Phenome Archive (EGA, https://www.ebi.ac.uk/ega/) under accession number EGAS00001002926. Plague reads were deposited in the European Nucleotide Archive (ENA) under accession number PRJEB25891.

Online content

Any Methods, including any statements of data availability and Nature Research reporting summaries, along with any additional references and Source Data files, are available in the online version of the paper at https://doi.org/10.1038/s41586-018-0094-2.

Received: 18 April 2017; Accepted: 3 April 2018;
Published online 9 May 2018.

1. Haak, W. et al. Massive migration from the steppe was a source for Indo-European languages in Europe. *Nature* **522**, 207–211 (2015).
2. Allentoft, M. E. et al. Population genomics of Bronze Age Eurasia. *Nature* **522**, 167–172 (2015).
3. Mathieson, I. et al. Genome-wide patterns of selection in 230 ancient Eurasians. *Nature* **528**, 499–503 (2015).
4. Chlenova, N. L. in *The Archaeology of the Steppes: Methods and Strategies* (ed. Genito, B.) 499–540 (Istituto Universitario Orientale, Naples, 1994).
5. Grakov, B. N., Yelagina, N. G. & Yatsenko, I. V. *The Early Iron Age* (Moscow State Univ. Press, Moscow, 1977).
6. Kristiansen, K. *Europe Before History* (Cambridge Univ. Press, Cambridge, 2000).
7. Parzinger, H. *Die Frühen Völker Eurasiens: vom Neolithikum bis zum Mittelalter* (CH Beck, München, 2006).
8. Alekseev, A. in *The Golden Deer of Eurasia: Scythian and Sarmatian Treasures from the Russian Steppes* (eds Aruz, J. et al.) 41–47 (The Metropolitan Museum of Art, New York, 2006).
9. Yablonsky, L. in *The Golden Deer of Eurasia: Scythian and Sarmatian Treasures from the Russian Steppes* (eds Aruz, J. et al.) 24–31 (The Metropolitan Museum of Art, New York, 2006).
10. Bashilov, V. A. & Yablonsky, L. T. in *Kurgans, Ritual Sites, and Settlements: Eurasian Bronze and Iron Age* (eds Davis-Kimball, J. et al.) 9–12 (Archaeopress, Oxford, 2000).
11. Unterländer, M. et al. Ancestry and demography and descendants of Iron Age nomads of the Eurasian Steppe. *Nat. Commun.* **8**, 14615 (2017).
12. Frachetti, M. D. Multiregional emergence of mobile pastoralism and nonuniform institutional complexity across Eurasia. *Curr. Anthropol.* **53**, 2–38 (2012).
13. Kohl, P. L. Shared social fields: evolutionary convergence in prehistory and contemporary practice. *Am. Anthropol.* **110**, 495–506 (2008).
14. Alexander, D. H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* **19**, 1655–1664 (2009).
15. Dybo, A. V. *Lingvističeskie kontakty rannix tjurkov. Leksičeskij fond* (Vostočnaja Literatura, Moscow, 2007).
16. Keyser-Tracqui, C., Crubézy, E. & Ludes, B. Nuclear and mitochondrial DNA analysis of a 2,000-year-old necropolis in the Egyin Gol Valley of Mongolia. *Am. J. Hum. Genet.* **73**, 247–260 (2003).
17. Keyser-Tracqui, C., Crubézy, E., Pamzsav, H., Varga, T. & Ludes, B. Population origins in Mongolia: genetic structure analysis of ancient and modern DNA. *Am. J. Phys. Anthropol.* **131**, 272–281 (2006).
18. Kim, K. et al. A western Eurasian male is found in 2000-year-old elite Xiongnu cemetery in Northeast Mongolia. *Am. J. Phys. Anthropol.* **142**, 429–440 (2010).
19. Pohl, W. in *A Companion to Ethnicity in the Ancient Mediterranean* (ed. McInerney, J.) 555–568 (Wiley Blackwell, Chichester, 2014).
20. De la Vaissière, É. Huns et Xiongnu. *Cent. Asiat. J.* **49**, 3–26 (2005).
21. Mallory, J. P. *In Search of the Indo-Europeans. Language, Archaeology and Myth* (Thames & Hudson, London, 1989).
22. Sinor, D. in *The Cambridge History of Early Inner Asia* (ed. Sinor, D.) 285–316 (Cambridge Univ. Press, Cambridge, 1990).
23. Findley, C. V. *The Turks in World History* (Oxford Univ. Press, Oxford, 2004).
24. Kradin, N. in *Xiongnu Archaeology: Multidisciplinary Perspectives of the First Steppe Empire in Inner Asia* (eds Brosseder, U. & Miller, B. K.) 77–96 (Universität Bonn, Bonn, 2011).
25. Golden, P. B. *An Introduction to the History of the Turkic Peoples: Ethnogenesis and State-Formation in Medieval and Early Modern Eurasia and the Middle East* (Harrassowitz, Wiesbaden, 1992).
26. Hildinger, E. *Warriors of the Steppe: a Military History of Central Asia, 500 BC to 1700 AD* (Da Capo Press, Cambridge, 1997).
27. Kradin, N. N. & Skrynnikova, T. D. *Imperiya Imperija Chingis Čingis-Khana Xana [The Genghis Khan Empire]* (Vostočnaja Literatura, Moscow, 2006).
28. Little, L. K. *Plague and the End of Antiquity: the Pandemic of 541–750* (Cambridge Univ. Press, Cambridge, 2007).
29. Wagner, D. M. et al. *Yersinia pestis* and the plague of Justinian 541–543 AD: a genomic analysis. *Lancet Infect. Dis.* **14**, 319–326 (2014).
30. Cui, Y. et al. Historical variations in mutation rate in an epidemic pathogen, *Yersinia pestis. Proc. Natl Acad. Sci. USA* **110**, 577–582 (2013).
31. Rasmussen, S. et al. Early divergent strains of *Yersinia pestis* in Eurasia 5,000 years ago. *Cell* **163**, 571–582 (2015).
32. Sun, Y.-C. C., Jarrett, C. O., Bosio, C. F. & Hinnebusch, B. J. Retracing the evolutionary path that led to flea-borne transmission of *Yersinia pestis. Cell Host Microbe* **15**, 571–582 (2015).
33. Kuz'mina, E. E. *The Origin of the Indo-Iranians* (Brill, Leiden, 2007).
34. Tremblay, X. Irano-Tocharica et Tocharo-Iranica. *Bull. Sch. Orient. Afr. Stud.* **68**, 421–449 (2005).
35. Nichols, J. in *Language Contact in Times of Globalization* (eds Hasselblatt, C. et al.) 177–195 (Rodopi, Amsterdam, 2011).
36. Johanson, L. in *The Turkic Languages* (eds Johanson, L. & Csató, É. Á.) 81–125 (Routledge, London, 1998).
37. Johanson, L. in *The Handbook of Language Contact* (ed. Hickey, R.) 652–672 (Wiley-Blackwell, Chichester, 2010).
38. Janhunen, J. *Manchuria: an Ethnic History* (The Finno-Ugrian Society, Helsinki, 1996).
39. Doerfer, G. *Türkische und Mongolische Elemente im Neupersischen 1–4* (Harrassowitz, Wiesbaden, 1963–1975).
40. Goldberg, A. et al. Ancient X chromosomes reveal contrasting sex bias in Neolithic and Bronze Age Eurasian migrations. *Proc. Natl Acad. Sci. USA* **114**, 2657–2662 (2017).

Acknowledgements We thank K. Magnussen, L. Petersen, C. Mortensen and A. Seguin-Orlando at the Danish National Sequencing Centre for producing the analysed sequences; P. Reimer and S. Hoper at the 14Chrono Center Belfast for providing accelerator mass spectrometry dating; S. Hackenbeck for discussing palaeodietary reconstructions; D. Christiansen Appelt, B. Heyerdahl, the Explico Foundation team, J. Isakova, B. Daulet, A. Tairov, N. Abduov, B. Tudiyarov, V. Volkov, M. Akhchurin, I. Baimukhan, N. Namdakov, Y. Yusupov, E. Ramankulov, A. Nurgaziyev and A. Kusaev for important assistance in fieldwork; J. Stenderup, P. V. Olsen and T. Brand for technical assistance in the laboratory; all involved archaeologists, historians and geographers from Kazakhstan: A. Suslov, I. Erofeeva, E. Nurmaganbetov, B. Kozhakhmetov, N. Loman, Y. Parshin, S. Ladunskiy, M. Bedelbaeva, A. Marcsik, O. Gábor, M. Půlpán, Y. Kubeev, R. Zhumashev, K. Omarov, S. Kasymov and U. Akimbayeva; P. Rodzianko for creating the initial contact between P.d.B.D., S.E. and E.U.; and S. Jacobsen and J. O'Brien for translating and proofreading Russian contributions. E.W. thanks St. John's College, Cambridge for support and for providing an environment facilitating scientific discussions. B.Boldg. thanks the Taylor Family-Asia Foundation Endowed Chair in Ecology and Conservation Biology. The project was funded by the Danish National Research Foundation (E.W.), the Lundbeck Foundation (E.W.) and KU2016 (E.W.).

Reviewer information *Nature* thanks T. Higham, D. Anthony, B. Shapiro, R. Dennell and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions E.W. initiated and led the study. P.d.B.D., E.W., E.U. and E.H. designed the study. P.d.B.D. and N.M. produced the data. P.d.B.D., N.M., S.R., M.S., G.R., T.Ko., A.Gol., M.W.P., A.G.P. and K.N. analysed or assisted in analysis of data. P.d.B.D., E.W. and K.K. interpreted results with considerable input from M.S., R.N., M.P., N.K., S.R., L.O., M.E.A. and J.V.M.-M. P.d.B.D., E.W., K.K., M.P. and S.R. wrote the manuscript with considerable input from N.K., L.H., M.S., R.N., M.E.A., L.O. and J.V.M.-M., with contributions from all authors. P.d.B.D., M.E.A., L.O., E.U., N.B., V.L., G.A., K.A., A.Ald., A.Alp., G.B., V.I.B., A.B., B.Boldb., B.Boldg., C.D., S.E., D.E., R.D., E.D., V.E., K.M.F., A.Gor., A.Gr., H.H., T.H., Z.K., R.K., E.K., A.Ko., T.Ku., A.Ku., I.K., N.L., A.M., V.K.M., I.V.M., I.M., E.M., V.M., G.M., B.N., Z.O., I.P., K.P., V.S., I.S., A.L., K.-G.S., T.S., K.T., A.T., T.T., D.V., L.Y., S.U., V.V., A.W. and E.H. excavated, curated, sampled and/or described analysed skeletons; all authors contributed to final interpretation of data.

Competing interests The authors declare no competing interests.

Additional information

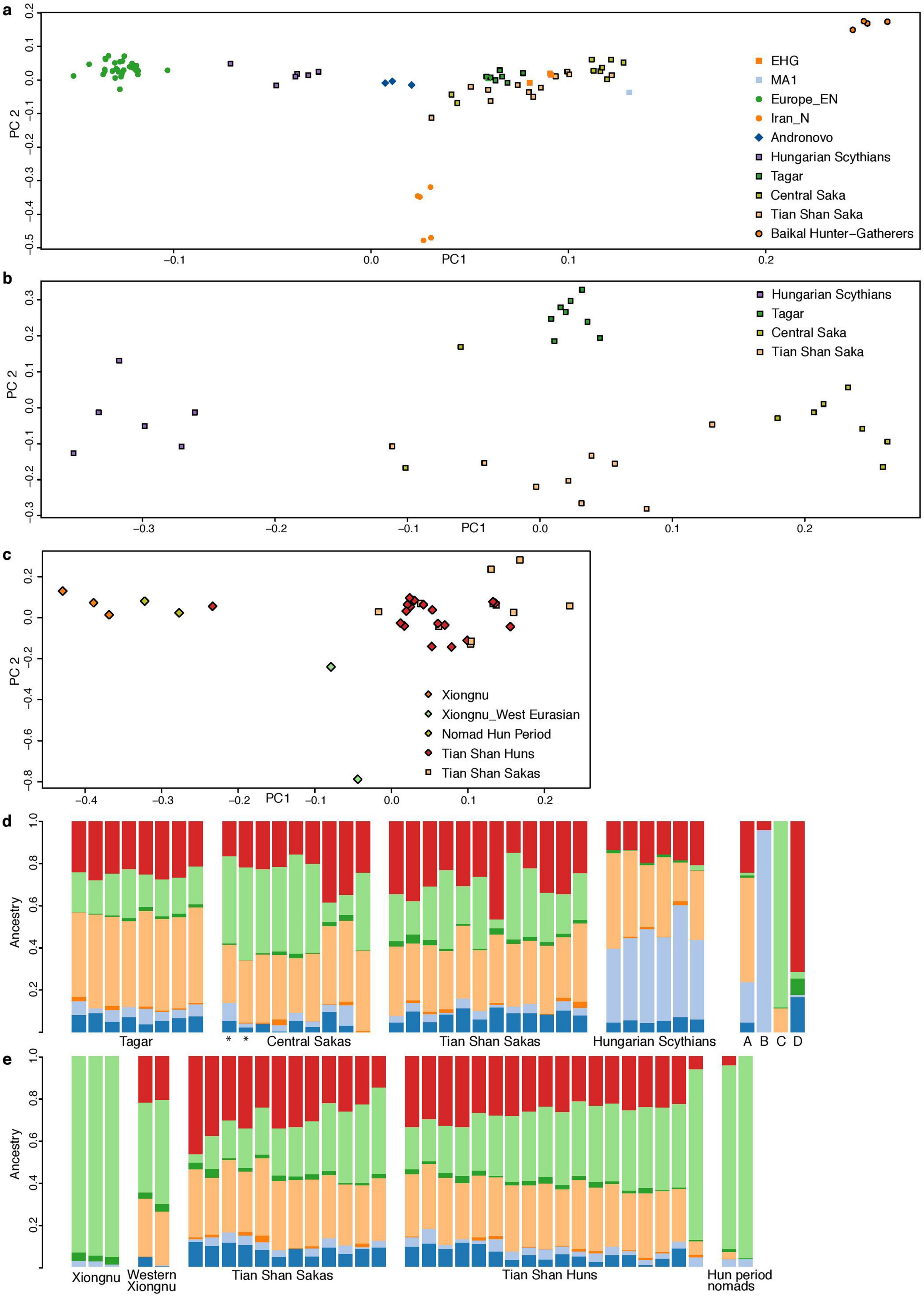
Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0094-2.

Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0094-2.

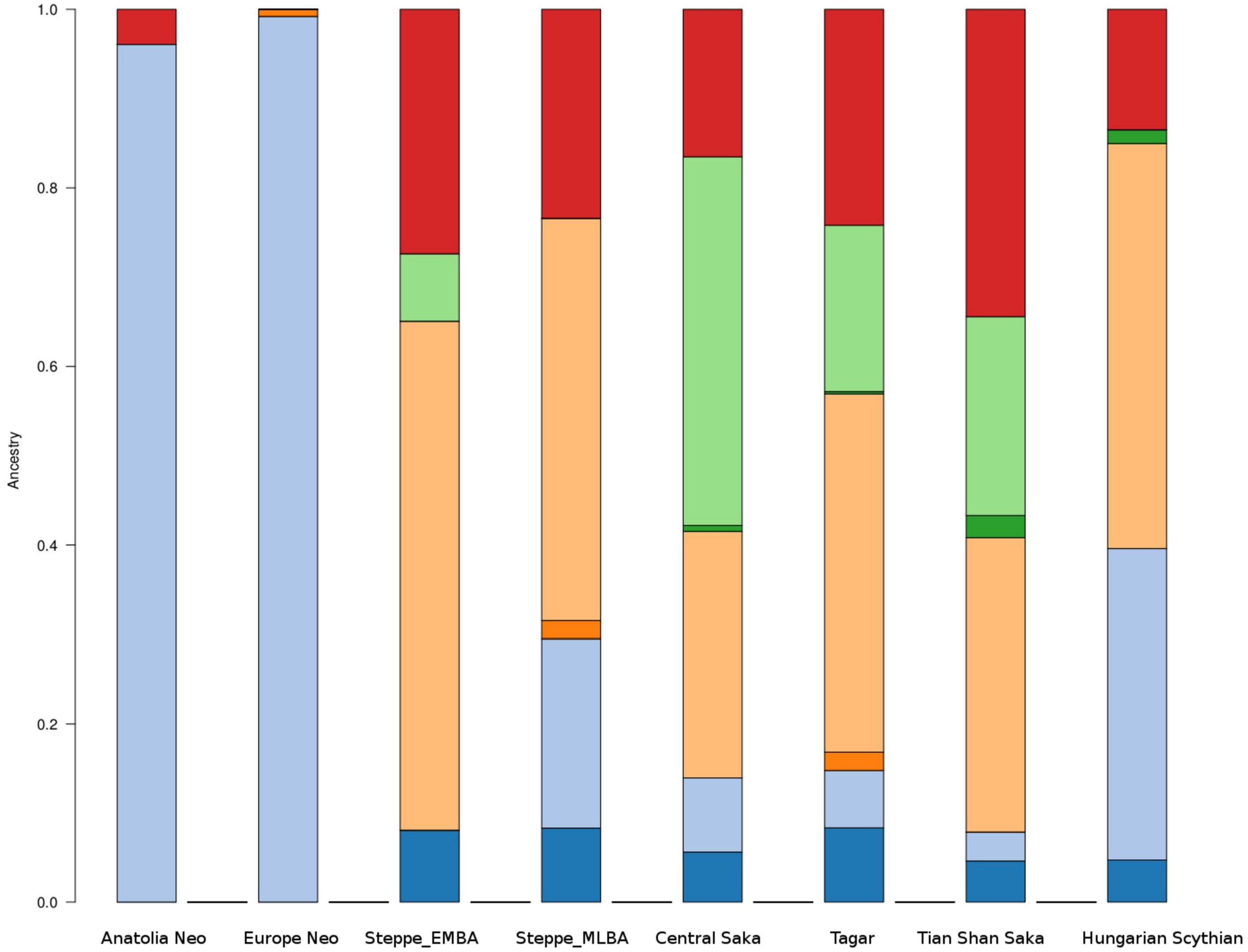
Reprints and permissions information is available at http://www.nature.com/reprints.

Correspondence and requests for materials should be addressed to E.W.

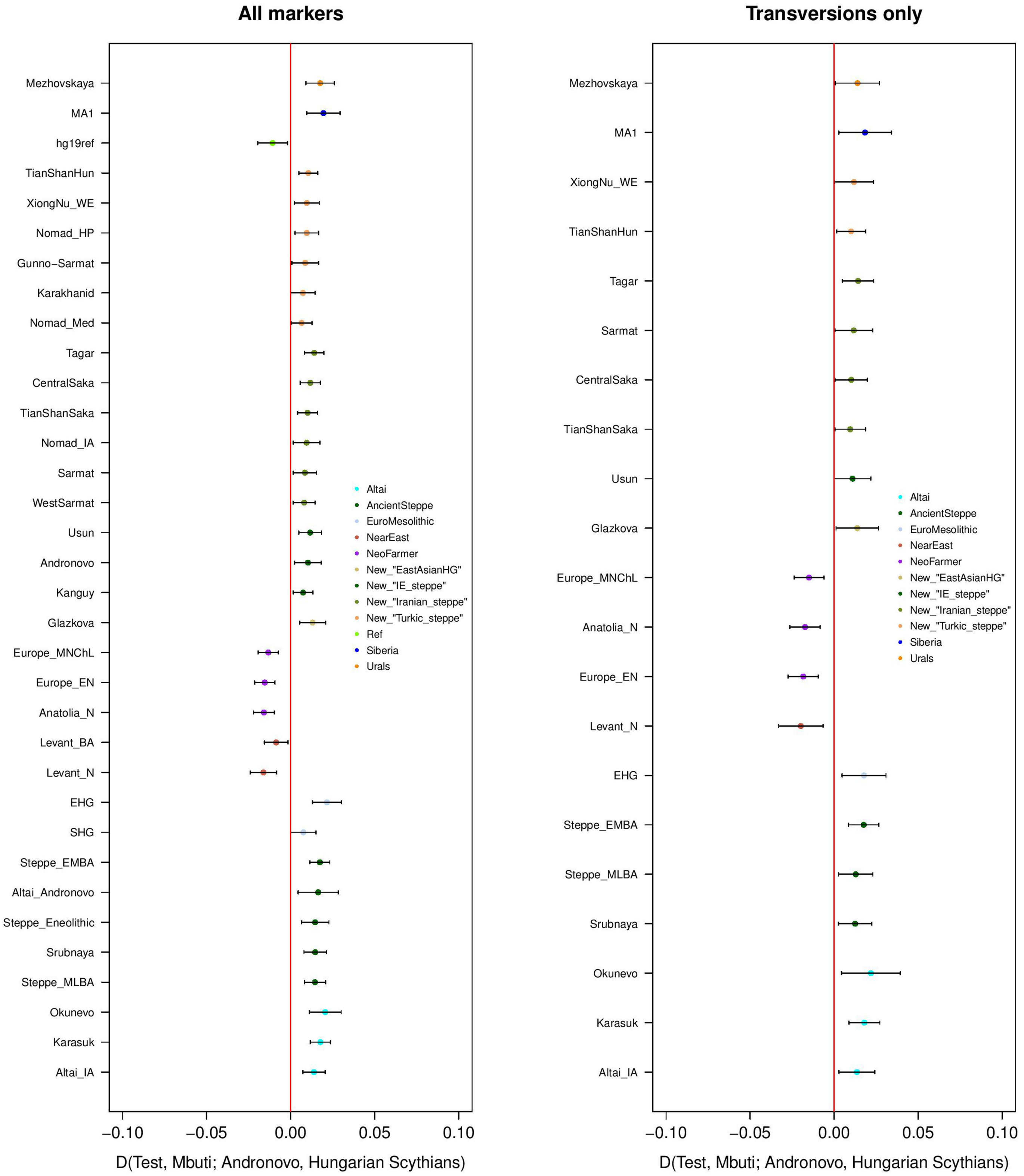
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



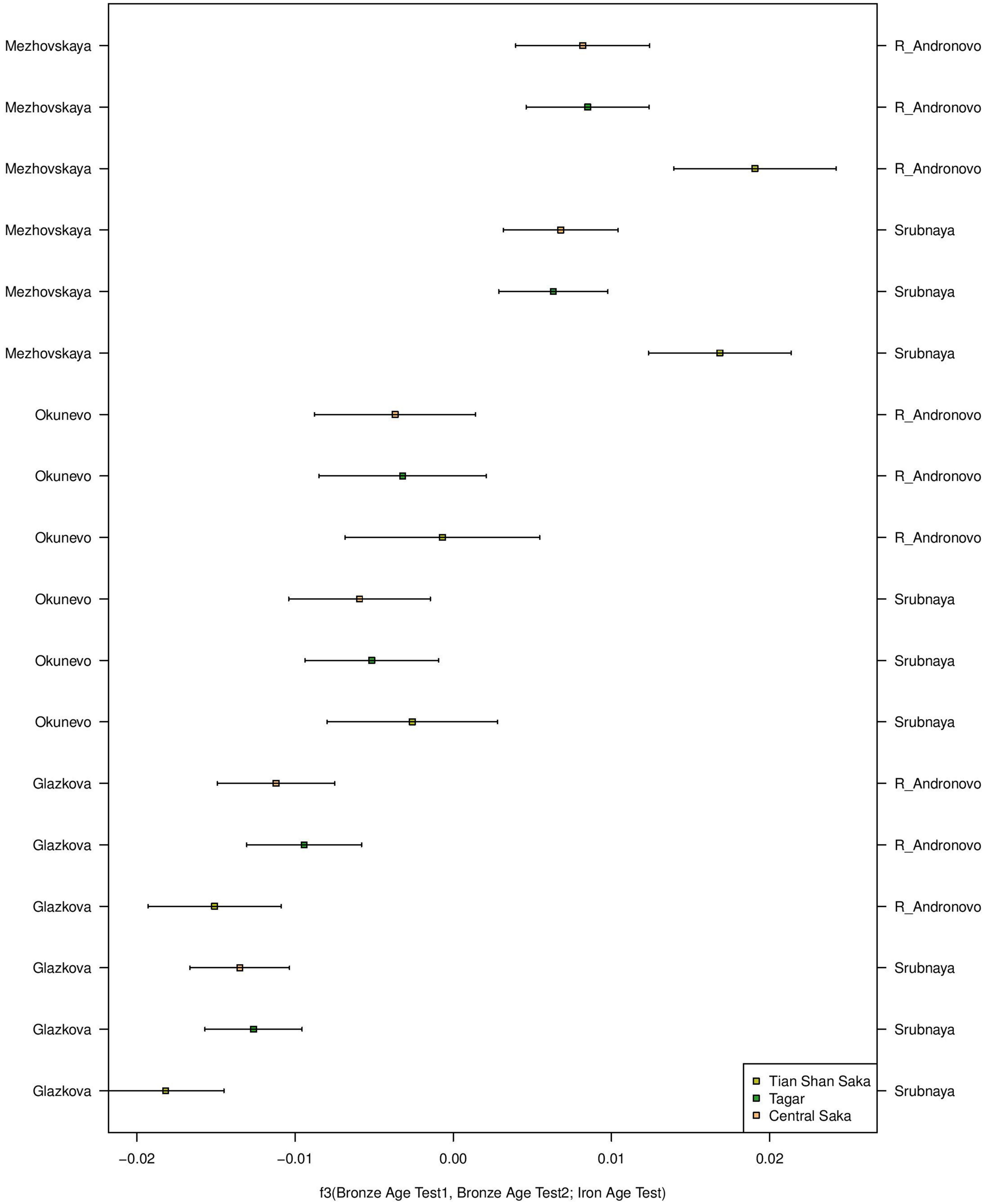
Extended Data Fig. 1 | Analyses of Iron Age clusters. **a**, PCA of Iron Age nomads and ancestral sources, explaining the diversity between them using 74 individuals at 242,406 autosomal single nucleotide polymorphism (SNP) positions. **b**, PCA of Iron Age nomads alone using 29 individuals at 242,406 autosomal SNP positions. **c**, PCA of Xiongnu, 'Western' Xiongnu, Tian Shan Huns, Nomads Hun Period, and Tian Shan Sakas, using 39 individuals at 242,406 autosomal SNP positions. **d**, Model-based clustering at $K = 7$ illustrating differences in ancestral proportions. Labelled individuals: A, Andronovo; B, Neolithic European (Europe_EN, in **a**); C, Baikal hunter-gatherers; D, Neolithic Iranian (Iran_N, in **a**). Here we illustrate the admixture analyses with $K = 7$ as it approximately identifies the major component of relevance (Anatolian/European farmer component, Caucasian ancestry, EHG-related ancestry and East Asian ancestry). The asterisk indicates an individual flagged as a genetic outlier. **d**, **e**, Results for model-based clustering analysis at $K = 7$. Here we illustrate the admixture analyses with $K = 7$ as it approximately identifies the major component of relevance (Anatolian/European farmer component, Caucasian ancestry, EHG-related ancestry and East Asian ancestry). Panel **d** is focused on the Iron Age, while **e** is focused on the transition to the Hun period.



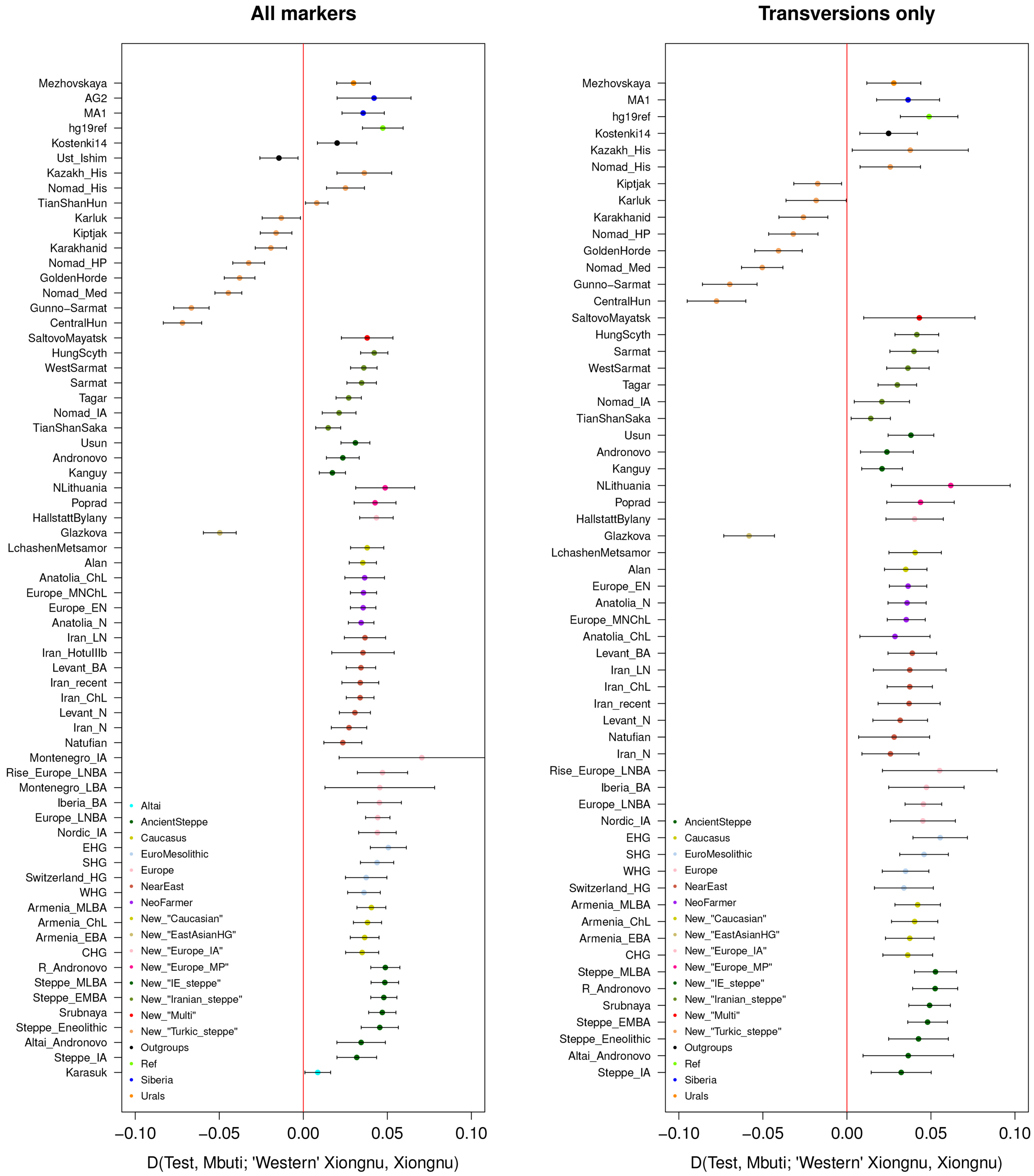
Extended Data Fig. 2 | Illustration of shared ancestry between Neolithic farmers and Iron Age nomads. Results for model-based clustering analysis at $K = 7$, plotting only one individual from relevant groups, to illustrate shared ancestry between Neolithic farmers from Europe, Late Bronze Age nomads and Iron Age nomads, not shared with Early Bronze Age nomads. MBLA, Middle-to-Late Bronze Age; Neo, Neolithic.



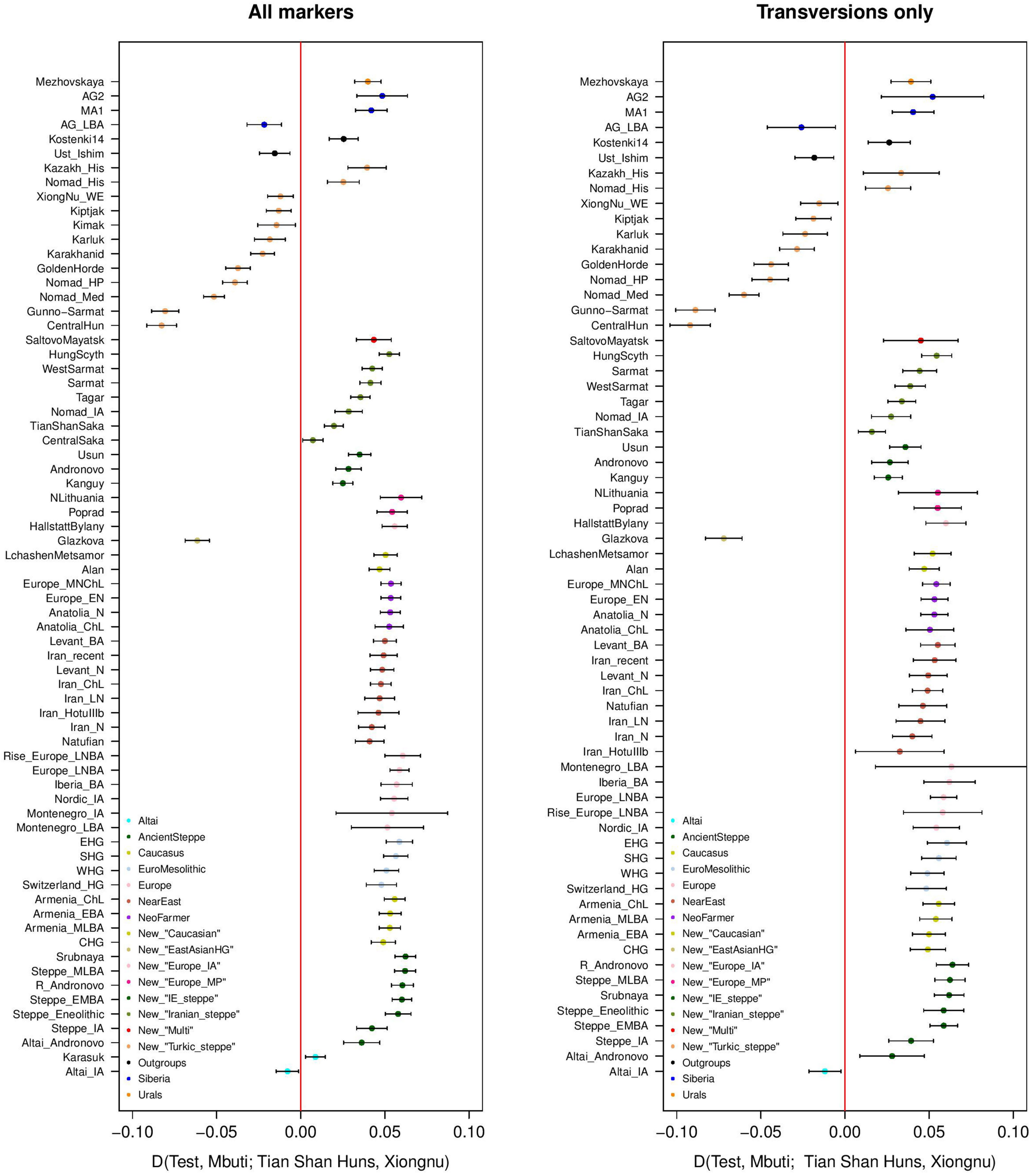
Extended Data Fig. 3 | Illustration of gene flow into Hungarian Scythians. We represent all *D*(Test, Mbuti; Andronovo, Hungarian Scythians) that deviate significantly from 0 (that is, higher than 3× the standard errors). The reported numbers are the *D*-statistics and the 3 standard errors were plotted as error bars. The number of individuals per population can be found in Supplementary Tables 3, 4.



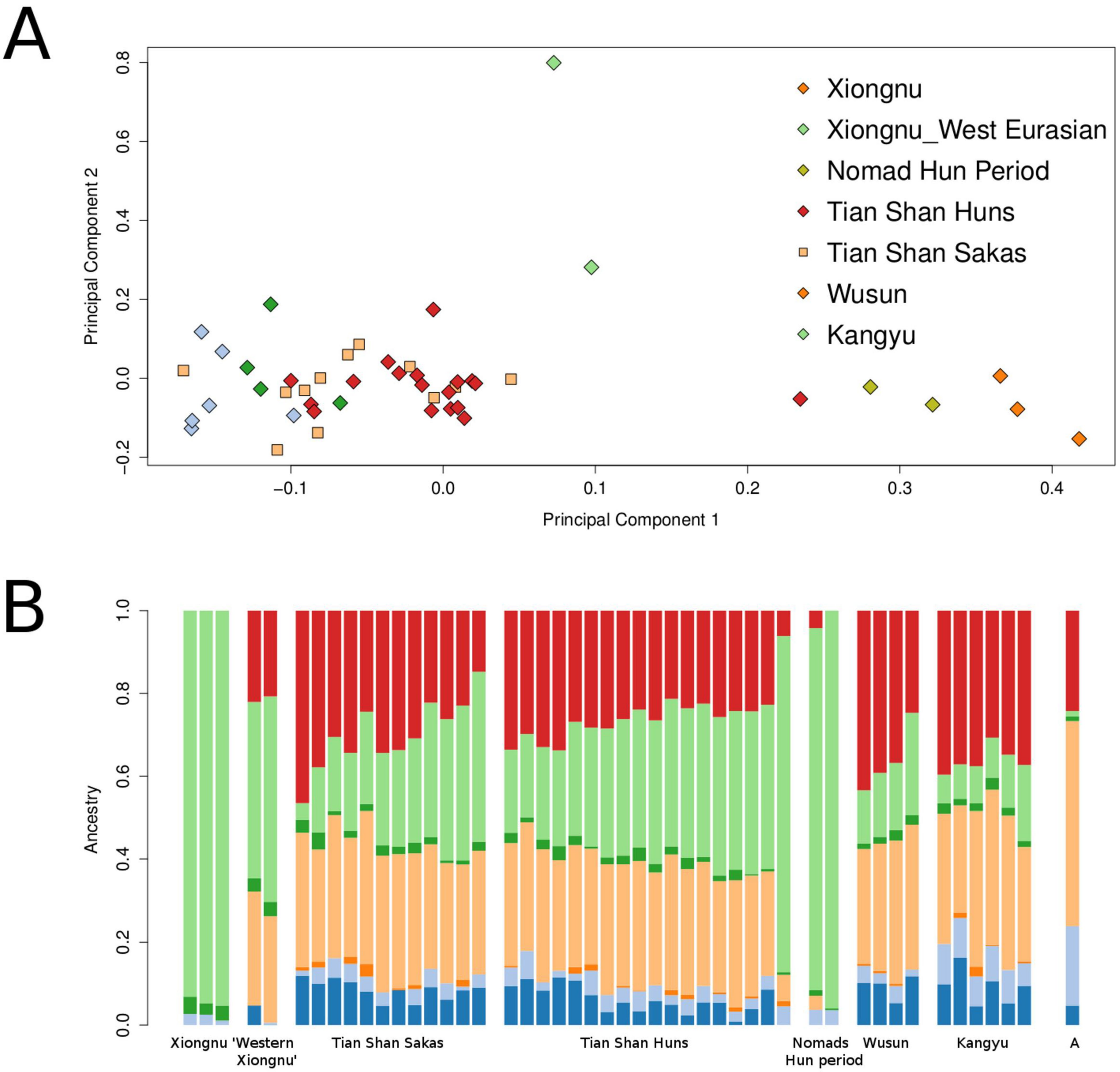
Extended Data Fig. 4 | Illustration of negative admixture f_3 statistics for Iron Age populations. Plot shows $f3$(Bronze Age Test 1, Bronze Age Test 2; Iron Age Test). The reported numbers are of the f_3 statistics, and the 3 standard errors were plotted as errors bars. The number of individuals per population can be found in Supplementary Table 3.



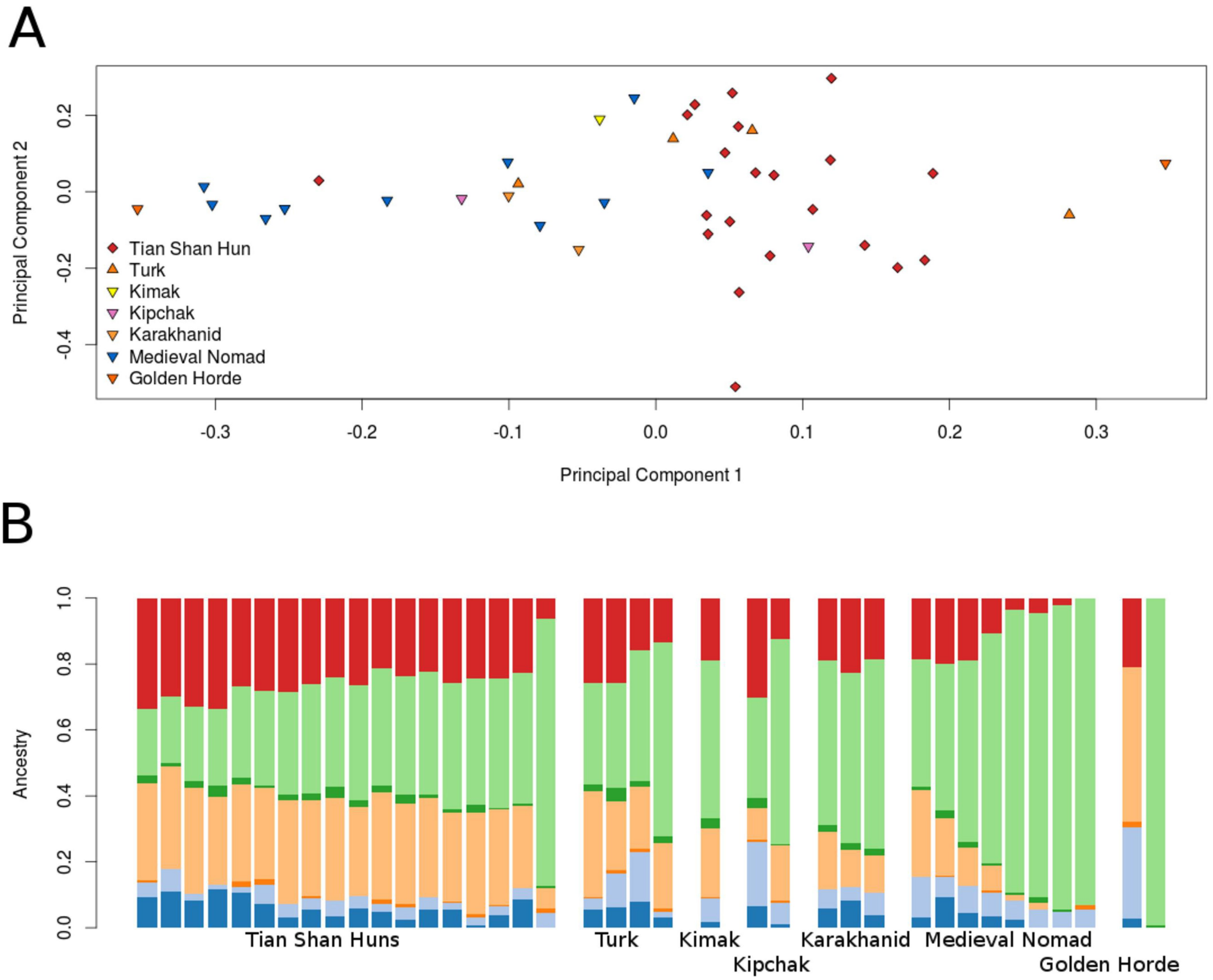
Extended Data Fig. 5 | Illustration of West Eurasian gene flow into groups forming the Xiongnu culture. We represent all *D*(Test, Mbuti; 'Western' Xiongnu, Xiongnu) that deviate significantly from 0 (that is, higher than 3× the standard errors). The reported numbers are the *D*-statistics and the 3 standard errors were plotted as error bars. The number of individuals per population can be found in Supplementary Tables 3, 4.



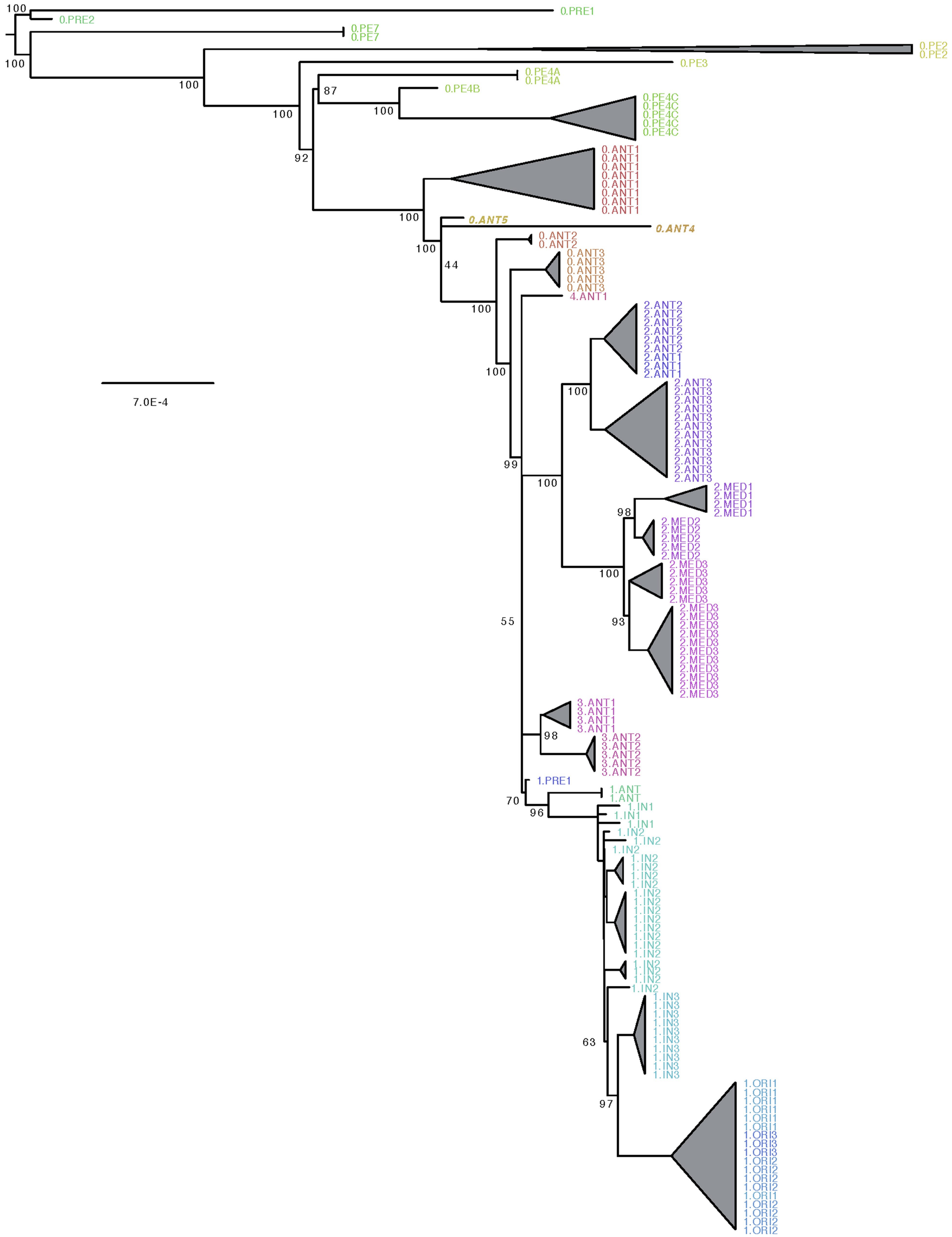
Extended Data Fig. 6 | Illustration of West Eurasian ancestry in early Tian Shan Huns. We represent all *D*(Test, Mbuti; Tian Shan Huns, Xiongnu) that deviate significantly from 0 (that is, higher than 3× the standard errors). The reported numbers are the *D*-statistics and the 3 standard errors were plotted as error bars. The number of individuals per population can be found in Supplementary Tables 3, 4.



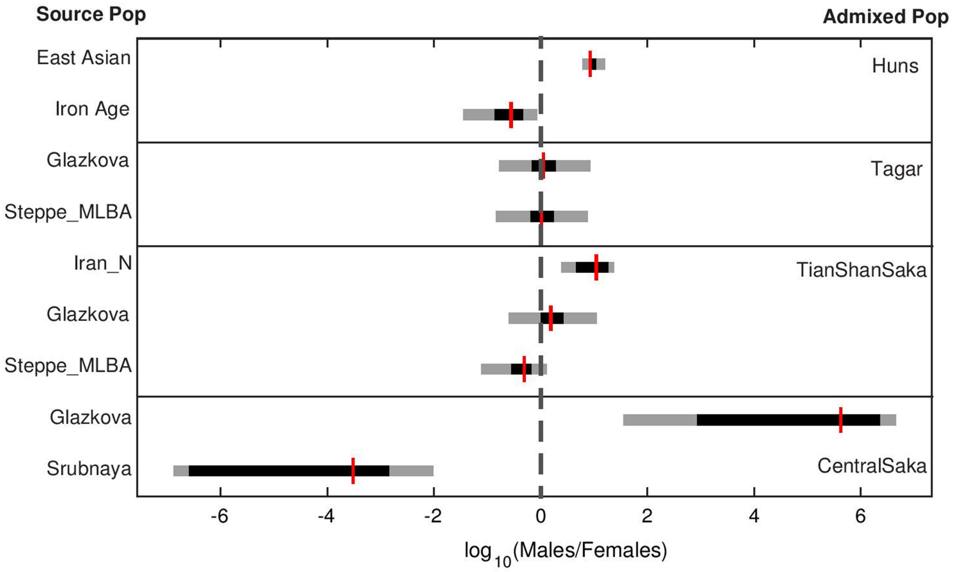
Extended Data Fig. 7 | Analyses of Xiongnu and Hun period population clusters. **a**, PCA of Xiongnu, 'Western' Xiongnu, Tian Shan Huns, Hun-period nomads, Tian Shan Sakas, Kangju and Wusun, including 49 individuals analysed at 242,406 autosomal SNP positions. **b**, Results for model-based clustering analysis at $K = 7$. Here we illustrate the admixture analyses with $K = 7$ as it approximately identifies the major component of relevance (Anatolian/European farmer component, Caucasian ancestry, EHG-related ancestry and East Asian ancestry). Individual A is a southern Siberian individual associated with the Andronovo culture.



Extended Data Fig. 8 | Analyses of Turk- and Medieval-period population clusters. a, PCA of Tian Shan Hun, Turk, Kimak, Kipchack, Karakhanid and Golden Horde, including 28 individuals analysed at 242,406 autosomal SNP positions. **b**, Results for model-based clustering analysis at $K = 7$. Here we illustrate the admixture analyses with $K = 7$ as it approximately identifies the major component of relevance (Anatolian/European farmer component, Caucasian ancestry, EHG-related ancestry and East Asian ancestry).



Extended Data Fig. 9 | Maximum likelihood phylogenetic reconstruction of *Y. pestis*. This tree reveals the basal position of the Tian Shan sample (0.ANT5, DA101, AD 186) compared to the Justinian plague sample (0.ANT4, A120, AD 536). These two samples are shown in orange italics. Other ancient plague samples included in the tree are Bronze Age samples (0.PRE1 and 0.PRE2) and a Black Death sample (1.PRE1). Numbers on nodes indicate bootstrap support (not all of which are shown, for clarity) and certain branches have been collapsed for clarity. Branch lengths are substitutions per site.



Extended Data Fig. 10 | Analyses of sex-specific contributions to Iron Age populations. Estimates of the male and female contributions from each source populations (left column) to each of the four admixed populations (right column) using a previously published method[40]. For each admixed population, we compared the observed mean autosomal and X-chromosomal ancestry, estimated in qpAdm, to that calculated under a constant admixture model on a grid of sex-specific contribution parameters ranging from 0 to 1 in 0.025 increments using a Euclidean distance. The logarithms of the ratio of male to female contribution parameters that produce the smallest 0.1% of distances from the data are plotted, with the full range of parameter values in grey, the middle 50% in black, and the median value in red. The dashed line indicates equal male and female contributions.

ARTICLE

https://doi.org/10.1038/s41586-018-0103-5

Diverse reprogramming codes for neuronal identity

Rachel Tsunemoto[1,2,8], Sohyon Lee[1,8], Attila Szűcs[3,4], Pavel Chubukov[1], Irina Sokolova[5], Joel W. Blanchard[1], Kevin T. Eade[1], Jacob Bruggemann[6], Chunlei Wu[6], Ali Torkamani[7], Pietro Paolo Sanna[5] & Kristin K. Baldwin[1,2]*

The transcriptional programs that establish neuronal identity evolved to produce the rich diversity of neuronal cell types that arise sequentially during development. Remarkably, transient expression of certain transcription factors can also endow non-neural cells with neuronal properties. The relationship between reprogramming factors and the transcriptional networks that produce neuronal identity and diversity remains largely unknown. Here, from a screen of 598 pairs of transcription factors, we identify 76 pairs of transcription factors that induce mouse fibroblasts to differentiate into cells with neuronal features. By comparing the transcriptomes of these induced neuronal cells (iN cells) with those of endogenous neurons, we define a 'core' cell-autonomous neuronal signature. The iN cells also exhibit diversity; each transcription factor pair produces iN cells with unique transcriptional patterns that can predict their pharmacological responses. By linking distinct transcription factor input 'codes' to defined transcriptional outputs, this study delineates cell-autonomous features of neuronal identity and diversity and expands the reprogramming toolbox to facilitate engineering of induced neurons with desired patterns of gene expression and related functional properties.

Neurons comprise a conspicuously diverse but clearly recognizable cell type. All neurons share defining features such as electrical excitability and synaptic connectivity. However, in even the simplest organisms, neurons also exhibit extensive diversity that affords each species its unique sensory modalities, behaviours and cognitive capabilities. The extent to which this diversity reflects the action of intrinsic cellular programs or depends on environmental and developmental cues is a central question in neuroscience.

Despite the elaborate sequential mechanisms that specify cell identity during development, recent studies have shown that transient overexpression of transcription factors can stably reprogram cells from one lineage to another without cell division, including the direct conversion of fibroblasts into iN cells using three transcription factors[1–3]. This discovery has enabled engineering of iN cells that resemble various endogenous subtypes, typically by adding transcription factors to the orginal neuron-inducing factors[3–10]. The majority of these protocols included achaete-scute homolog 1 (ASCL1, encoded by the *Ascl1* gene), suggesting that this may be an essential factor[11]. However, we showed that replacing ASCL1 with neurogenin 1 (encoded by *Neurog1*) or neurogenin 2 (encoded by *Neurog2*), while co-expressing POU4F1 (also known as BRN3A, encoded by *Pou4f1*), selectively induced a population of cells that resembled dorsal root ganglion (DRG) sensory neurons[12].

These studies raised several important questions. First, is the capacity to ectopically induce neuronal identity limited to only a few sets of transcription factors or might there be a larger set of inducing factors? Second, what features of neuronal identity and diversity can be produced outside the context of development and the brain? Third, how might we generate iN cells with desired functional properties or patterns of gene expression?

To explore these questions, we screened 598 pairs of transcription factors, testing their ability to induce neuronal identity in fibroblasts. Unexpectedly, more than 12% (76 of 598) of the transcription factor pairs could reprogram fibroblasts into iN cells that express key neuronal markers, exhibit neuronal morphologies, are electrically active and can form synaptic connections without co-culturing with glia. Neurons within a given iN cell population are relatively homogenous, while different iN cell populations exhibit transcriptional and functional diversity. These studies define a new set of cell-autonomous transcriptional networks underlying neuronal identity and establish a database of transcription factor 'codes' to produce iN cells that express desired receptors, neurotransmitters, ion channels, synaptic proteins and other useful features of neuronal diversity.

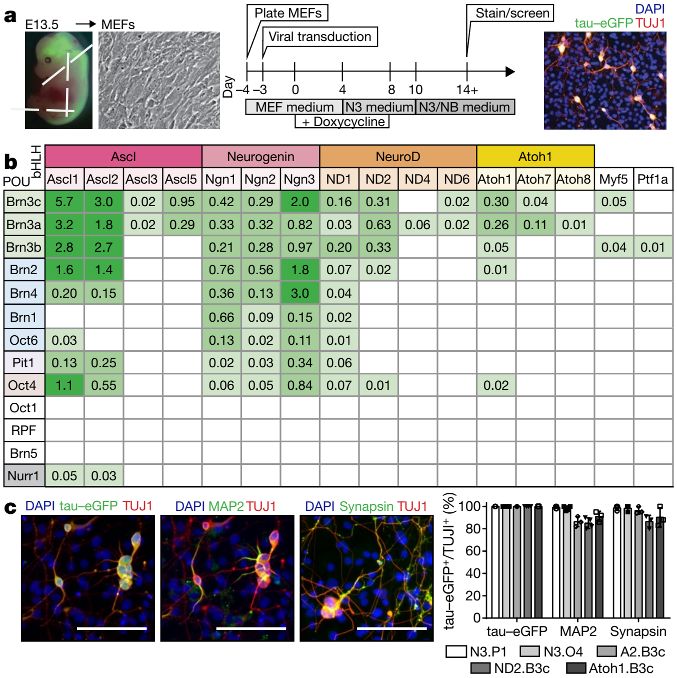
An unbiased screen of transcription factor pairs

We generated an inducible library of 598 transcription factor pairs comprised of 46 basic-helix–loop–helix (bHLH), one nuclear receptor and 12 Pit-Oct-Unc (POU) transcription factors, by cloning cDNAs encoding the transcription factors into doxycycline-inducible lentiviral vectors. Two weeks after induction of mouse embryonic fibroblasts (MEFs), 76 (12.7%) of the 598 transcription factor pairs produced cells that expressed the neuron-specific class III β-tubulin (TUJ1, encoded by *Tubb3*) and had neuronal morphologies (Fig. 1a, b, Extended Data Fig. 1a, Supplementary Table 1). Most individual transcription factors had no effect; however, four factors produced rare TUJ1$^+$ cells with atypical morphologies—these were subtracted when scoring hits (Extended Data Fig. 1b, c). The majority of TUJ1$^+$ candidate iN cells also expressed the neuronal markers MAP2 (85–99%), synapsin (86–98%) and tau (based on later experiments using tau–eGFP knock-in mice) (Fig. 1c, d). All tested pairs (12 of 12) were also able to reprogram adult tail-tip fibroblasts (TTFs) and MEFs depleted of p75-expressing neural crest cells[12] (Extended Data Fig. 1d–h).

Neurons are most stringently defined by their electrophysiological properties. We analysed the electrophysiology of five candidate iN cell

---

[1]Department of Molecular and Cellular Neuroscience, Dorris Neuroscience Center, The Scripps Research Institute, La Jolla, CA, USA. [2]Neuroscience Graduate Program, University of California San Diego, La Jolla, CA, USA. [3]BioCircuits Institute, University of California San Diego, La Jolla, CA, USA. [4]MTA-ELTE-NAP B Neuronal Cell Biology Research Group, Eotvos Lorand University, Budapest, Hungary. [5]Molecular and Integrative Neurosciences Department, The Scripps Research Institute, La Jolla, CA, USA. [6]Integrative Structural and Computational Biology, The Scripps Research Institute, La Jolla, CA, USA. [7]Scripps Translational Science Institute, Scripps Health and The Scripps Research Institute, La Jolla, CA, USA. [8]These authors contributed equally: Rachel Tsunemoto, Sohyon Lee. *e-mail: kbaldwin@scripps.edu



POU \ bHLH	Ascl				Neurogenin			NeuroD				Atoh1				
	Ascl1	Ascl2	Ascl3	Ascl5	Ngn1	Ngn2	Ngn3	ND1	ND2	ND4	ND6	Atoh1	Atoh7	Atoh8	Myf5	Ptf1a
Brn3c	5.7	3.0	0.02	0.95	0.42	0.29	2.0	0.16	0.31		0.02	0.30	0.04		0.05	
Brn3a	3.2	1.8	0.02	0.29	0.33	0.32	0.82	0.03	0.63	0.06	0.02	0.26	0.11	0.01		
Brn3b	2.8	2.7			0.21	0.28	0.97	0.20	0.33			0.05			0.04	0.01
Brn2	1.6	1.4			0.76	0.56	1.8	0.07	0.02			0.01				
Brn4	0.20	0.15			0.36	0.13	3.0	0.04								
Brn1					0.66	0.09	0.15	0.02								
Oct6	0.03				0.13	0.02	0.11	0.01								
Pit1	0.13	0.25			0.02	0.03	0.34	0.06								
Oct4	1.1	0.55			0.06	0.05	0.84	0.07	0.01			0.02				
Oct1																
RPF																
Brn5																
Nurr1	0.05	0.03														

Fig. 1 | Screening transcription factor pairs for iN cell induction. **a**, Reprogramming schematic. Candidate iN cells were identified by expression of neuronal markers and neuronal morphologies. Red, TUJ1; green, tau–eGFP; blue, DAPI. **b**, Positive (76) and negative (522) transcription factor pairs identified in the screen. All tested POU and nuclear receptor factors are included but the 30 bHLH factors that did not produce iN cells are omitted for clarity. Each box contains the normalized average percentage of TUJ1$^+$ cells per well ($n = 3$ wells, 2×10^4 fibroblasts per well). **c**, MEFs were transfected with vectors encoding *Neurog3/Pou1f1* to generate iN cells. Immunofluorescence showing co-labelling of TUJ1$^+$ (red) candidate iN cells with tau–eGFP (green), MAP2 (green) and synapsin (green) with nuclei in blue (DAPI) from $n = 5$, 5 and 3 independent experiments, left to right, respectively. Scale bars, 100 μm. **d**, Percentage of TUJ1$^+$ cells that co-express tau–eGFP ($n = 574$), MAP2 ($n = 574$) or synapsin ($n = 293$) for iN cells induced by *Neurog3/Pou1f1* (N3.P1, $n = 5$, 5 and 3 independent experiments, respectively), *Neurog3/Pou5f1* (N3.O4, $n = 4$, 4 and 3 independent experiments, respectively), *Ascl2/Pou4f3* (A2.B3c, $n = 3$, 3 and 3 independent experiments, respectively), *Neurod2/Pou4f3* (ND2.B3c, $n = 4$, 4 and 3 independent experiments, respectively) and *Atoh1/Pou4f3* (Atoh1.B3c, $n = 3$, 3 and 3 independent experiments, respectively). *Pou4f3* is also known as *Brn3c*. Data are mean ± s.d.

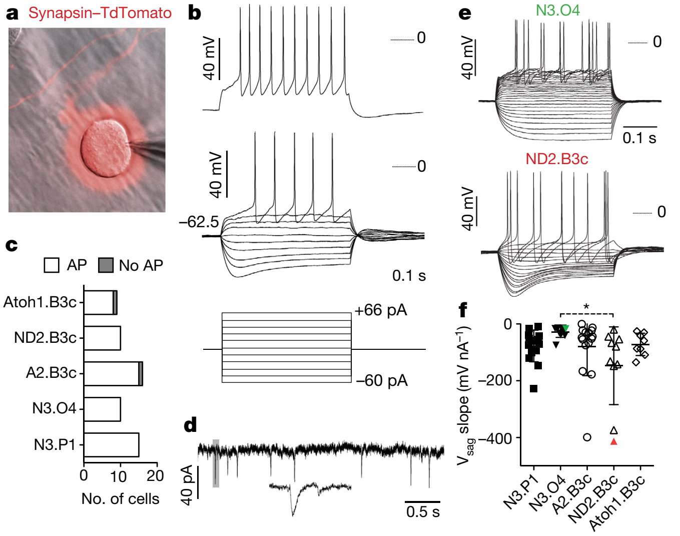


Fig. 2 | Electrophysiological properties of iN cells. **a**, Representative whole-cell patch-clamped candidate iN cell expressing synapsin–TdTomato (red). TdTomato intensity was adjusted to show neurites and soma. Scale bar, 25 μm. **b**, Membrane voltage responses from an iN cell generated with *Ascl2/Pou4f3* under whole-cell patch-clamp conditions at maximum current injection (top) and current steps until the first induction of action potentials (middle), with current traces (bottom). **c**, iN cells generated with five transcription factor pairs exhibit current-induced action potentials in the majority of cells: *Neurog3/Pou1f1* (N3.P1, 15 of 15 cells), *Neurog3/Pou5f1* (N3.O4; 10 of 10 cells), *Ascl2/Pou4f3* (A2.B3c; 15 of 16 cells), *Neurod2/Pou4f3* (ND2.B3c; 10 of 10 cells) and *Atoh1/Pou4f3* (Atoh1.B3c; 8 of 9 cells). AP, action potential. **d**, Current trace showing EPSCs from an iN cell generated with *Neurog3/Pou5f1*. **e**, Membrane voltage responses to depolarizing current steps of iN cells with neuronal morphology generated from *Neurog3/Pou5f1* (N3.O4, top) and *Neurod2/Pou4f3* (ND2.B3c, bottom). **f**, Quantification of voltage sag (V_{sag}) behaviour for candidate iN cells that exhibited current-induced action potentials: N3.P1 ($n = 15$ cells), N3.O4 ($n = 10$), A2.B3c ($n = 15$), ND2.B3c ($n = 10$) and Atoh1.B3c ($n = 8$). Voltage sag is plotted as the slope of the voltage sag versus current. Coloured points correspond to the plotted cells. Data are mean ± s.d., $^*P = 0.0207$, one-way ANOVA, Tukey's multiple comparison test.

populations that included transcription factors not previously reported to induce neurons in vitro (Fig. 2). The majority of candidate iN cells (expressing tau–eGFP and TdTomato driven by a synapsin promoter) (58 of 60, 97%) fired action potentials and displayed resting membrane potentials (−61.7 ± 7.8 mV) and other properties consistent with neuronal identity, whereas negative control cells did not. Voltage sag and input resistance were similar to those of endogenous neurons, but varied among some populations of iN cells (Fig. 2, Extended Data Fig. 2). Unexpectedly, we detected excitatory post-synaptic currents (EPSCs), in five recorded candidate iN cells (Fig. 2d, Extended Data Fig. 2n). The formation of active synapses is rarely observed in stem-cell derived neurons after 16–24 days of culture in the absence of glia. These results provide strong evidence that the candidate iN cells have acquired neuronal identity.

Both MEFs and human embryonic fibroblast-like cells (HEFs) derived from iPSCs can be reprogrammed with pairs of mouse transcription factors[12–14]. Here we show that mouse and human versions of NEUROG3 and POU1F1 (also known as PIT1), can reprogram HEFs with similar efficiency. All tested mouse transcription factor pairs (14 of 14) generated human iN cells that expressed TUJ1 and MAP2. The iN cells could fire action potentials ($n = 21$ of 27 iN cells produced with NEUROG3 and POU1F1, 77%) and exhibited membrane properties and voltage-dependent Na$^+$ and K$^+$ currents comparable to those reported previously[7–9,12] (Extended Data Fig. 3). These experiments establish the likely utility of the mouse screen for guiding human cell reprogramming experiments.

iN cells resemble endogenous neurons

The iN cells produced in this study exhibit neuronal morphologies, express defining markers of mature neurons and fire action potentials. To establish the extent to which the transcriptomes of iN cells resemble endogenous neural populations, we performed fluorescence-activated cell sorting (FACS) and RNA sequencing (RNA-seq) analysis on 35 mouse iN cell populations generated from MEFs derived from tau–eGFP knock-in mice, which express GFP specifically in neurons[3]. For comparison, we transcriptionally profiled whole brain RNA, control MEFs and eight FACS-purified populations of endogenous neurons, selected to encompass the peripheral and central nervous system and to include multiple neurotransmitter identities (Fig. 3a, Extended Data Fig. 4, Supplementary Table 2).

Principal component analysis (PCA) of RNA-seq data showed that the iN cell, endogenous neuron and brain populations intersected and were segregated from the MEFs (Fig. 3b). Using DESeq2[15], we identified the 3,860 genes that were upregulated in the iN cells in comparison to the MEFs. These genes were enriched for gene ontology (GO) terms associated with neuronal development, neuronal function and synaptic transmission[16]. Similarly, the 3,467 genes that were downregulated in

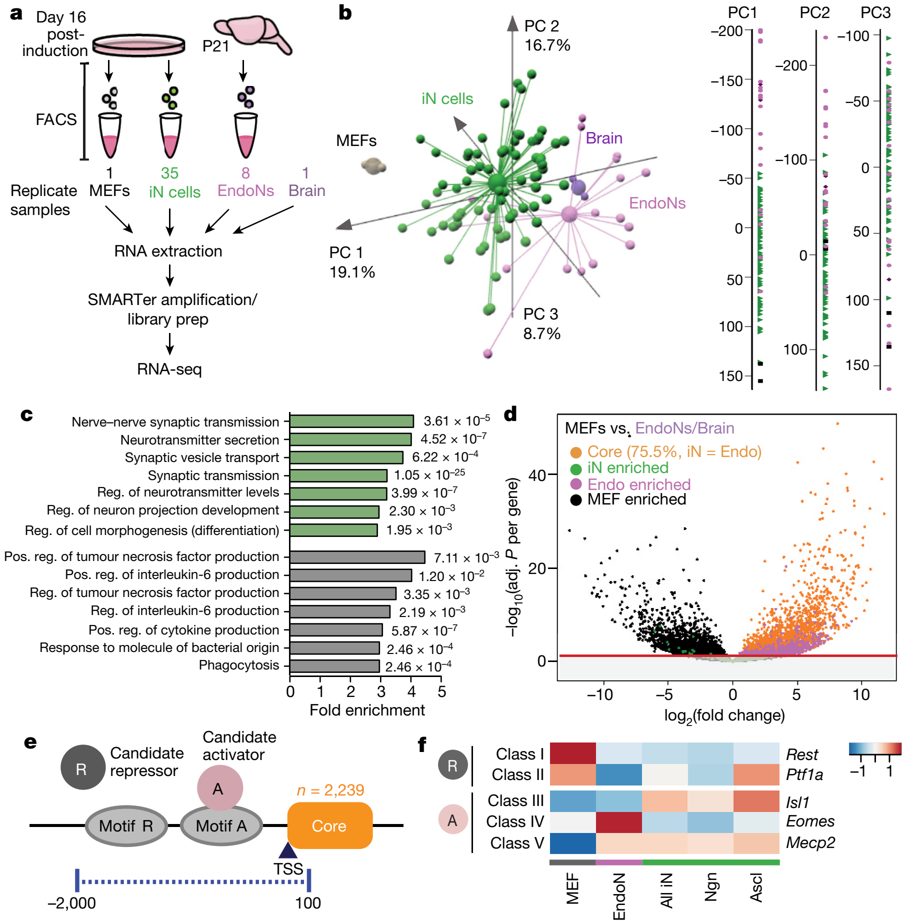


Fig. 3 | RNA-seq of iN cell populations. a, RNA-seq schematic for MEFs, induced neurons (iN cells), endogenous neurons (EndoNs) and whole brain samples. **b**, PCA from MEF ($n = 1$ population in duplicate biological samples, grey), iN cell ($n = 35$ in duplicate; green), endogenous neuron ($n = 6$ in duplicate, $n = 2$ in triplicate; purple) and brain ($n = 1$ in duplicate; dark purple) RNA-seq data. Loadings for principal components PC1, PC2 and PC3 are reported on x, y and z axes, respectively. Larger spheres represent centroids of each set of samples. Right, candidate iN cell populations (green), MEFs (grey), endogenous neurons (purple) and brain samples (dark purple) plotted along a vertical axis (loadings) for the first three principal components. **c**, Enriched GO terms for the differentially expressed genes based on DESeq2 comparisons of the MEFs compared to the iN cells plotted at fold enrichment with the associated P value at the end of each bar. Pos., position; reg., regulation. **d**, Abbreviated volcano plot of $\log_2$(fold change) versus $-\log$(adjusted P value per gene) for MEFs (black) versus the pooled endogenous neuron and brain RNA-seq data. Plotted are enriched core genes that are shared between iN cell and endogenous neurons and brain populations (orange, 75.5% of the significantly enriched endogenous neurons and brain genes), genes enriched in endogenous neurons and brain (purple, endo-enriched) and genes enriched in iN cells (green, iN cell-enriched). The red line represents $-\log(0.05$ P-adjusted value) (See Extended Data Fig. 5a). **e**, Schematic of predicted transcriptional repressors (R) and activators (A) that silence and activate neuronal genes in MEFs and neuronal populations, respectively. TSS, transcription start site. **f**, Expression of representative predicted transcriptional core gene regulators from each class identified by both HOMER and IPA plotted as DESeq2 vsd-normalized RNA-seq counts with groups averaged and scaled by row.

iN cells were enriched for GO terms associated with immune function and cell division (Fig. 3c, Supplementary Table 3). These analyses indicate that the iN cells have acquired global transcriptional programs similar to those of endogenous neurons.

Defining a core neuronal transcriptome

Gene networks shared by the iN cells and endogenous neuron and brain samples are of interest as they could define a 'core' neuronal transcriptome that arises both in vivo and in vitro. Conversely, genes found in the endogenous neuron and brain samples but not in the iN cells could reflect differences in cell types, neuronal maturity, the influence of exogenous signalling, or signs of incomplete reprogramming. In our analysis, the number of genes differentially expressed in the endogenous neuron and brain samples compared to MEFs ($n = 2,965$) was similar to that of the iN cells. We defined a candidate core neuronal transcriptome from the intersection (75.5% overlap) of the iN cell-enriched genes (versus MEFs) and the endogenous neuron and brain-enriched genes (versus MEFs). The number of genes shared between the core transcriptome and each individual population was similar among endogenous and iN cell populations (78 ± 7%, range 63–86%). Genes that were not enriched in iN cells were associated with GO terms related to glia, myelination and neural development, suggesting that iN cells express the majority of 'pan-neuronal' genes but not genes characteristic of glia or neuronal precursors (Fig. 3d, Extended Data Fig. 5a, b, Supplementary Table 3).

This highlights the capacity for diverse transcription factor pairs to independently converge on shared transcriptional signatures of neuronal identity and suggests that the result of direct reprogramming is similar to the result of normal neuronal development with respect to this core transcriptome.

This observation raises the question of whether the gene networks shared by iN cells and endogenous neurons are governed by similar mechanisms. Regulatory mechanisms resulting in neuronal identity may include loss of repressors of neuronal identity and/or gain of activators of neuronal genes (Fig. 3e). To identify candidate regulatory factors we applied two complementary bioinformatic tools, ingenuity pathway analysis (IPA) and hypergeometric optimization of motif enrichment (HOMER). IPA identified 39 candidate transcriptional

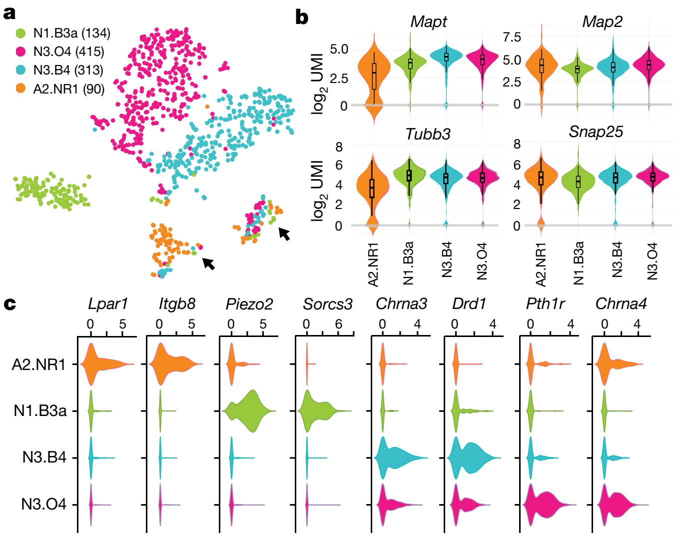


Fig. 4 | Single-cell RNA-seq of four iN cell populations. a–c, Cells are colour-coded as *Neurog1/Pou4f1* (N1.B3a, green, $n = 134$ cells), *Neurog3/Pou5f1* (N3.O4, pink, $n = 415$), *Neurog3/Pou3f4* (N3.B4, blue, $n = 313$) and *Ascl2/Nr4a2* (A2.NR1, orange, $n = 90$). **a**, *t*-SNE projection of 952 single cells derived from the four pairwise combinations of transcription factors. Black arrows indicate clusters of cells with low total UMI counts (Extended Data Fig. 5a). **b**, Single-cell $\log_2$ UMI expression of pan-neuronal (*Mapt*, *Tubb3*) and mature neuron (*Map2*, *Snap25*) markers. The rectangle of the box plot spans the interquartile range, the dividing segment denotes the median. The whiskers above and below the box represent the minimum and maximum. **c**, Single-cell $\log_2$ UMI expression of receptors and transmembrane proteins among the top 20 differentially expressed genes with each transcription factor combination relative to all other combinations (see Extended Data Fig. 5f). *Pou3f4* is also known as *Brn4*.

regulators that we divided into five classes (I–V). HOMER motif analyses identified 48 enriched motifs of which four overlapped with the results from IPA. Class I and II genes appear to be repressors because they are more highly expressed in MEFs compared to iN cells or endogenous neurons (Fig. 3f). We termed Class I genes 'common repressors' because they are expressed at lower levels in all iN cells and endogenous neurons, whereas Class II repressors exhibit subclass-specific patterns. The most enriched binding domain is that of the Class I gene *Rest*[17], which was recently shown to increase the efficiency of direct reprogramming when knocked down using silencing short hairpin RNAs (shRNAs), highlighting the potential for network analyses to uncover mechanisms that may enable improved reprogramming methods[18] (Fig. 3f, Extended Data Fig. 5c, d, Supplementary Table 3).

Class III and IV genes represent candidate activators, based on reduced expression in MEFs in comparison to endogenous neurons (Class III) or iN cells (Class IV). These include genes for transcription factors that are known to regulate neuronal differentiation and diversity, such as *Isl1* and *Eomes*. We identified only one candidate Class V 'common activator' transcription factor (*Mecp2*) that was expressed in all iN cells and endogenous neurons at higher levels than in MEFs, consistent with its high expression level in the brain and role in synaptic function. Together, these analyses suggest a model in which common mechanisms for gene derepression operate both in development and direct reprogramming, whereas gene activation may arise through different mechanisms (Fig. 3e, f, Extended Data Fig. 5c, d, Supplementary Table 3).

Single-cell RNA-seq of iN cells

A key question regarding the iN cell populations is the extent to which they exhibit homogeneity versus heterogeneity. We investigated this using single-cell RNA-seq (scRNA-seq). We analysed four representative iN cell populations (Fig. 4a, Methods). iN cells that were produced with the same transcription factor pair display relative homogeneity and generally cluster together. Cells falling outside of the main clusters have low total unique molecular identifier (UMI) counts, which is likely to lead to low expression of subtype genes and miscategorization (Fig. 4a, Extended Data Fig. 6a). The majority of iN cells expressed pan-neuronal markers such as *Mapt* (encoding tau) and *Tubb3* and mature markers (*Map2* and *Snap25*), although iN cells derived with the non-POU factor encoded by *Nr4a2* (also known as *Nurr1*) exhibited a slightly more graded expression, perhaps due to lower overall UMI counts (Fig. 4b).

Population RNA-seq experiments detected low levels of myocyte-related genes in iN cells produced using an Ascl family member, but not in other populations. Mapping these genes onto the *Ascl2/Nr4a2* single-cell data identified a small subpopulation of cells (3 of 90) with that exhibited coordinated co-expression of these genes, consistent with previous studies[19] (Extended Data Figs. 6b, 7b).

Similarly, the weak residual expression of MEF genes observed in population RNA-seq of iN cells could derive from rare contaminating MEFs in the sorted cells and/or residual MEF gene expression in the iN cells. We sorted and profiled single MEFs (tau–eGFP^{-}) and iN cells (tau–eGFP^{+}) from the same reprogramming experiment. This revealed a small population of cells with strong MEF gene expression, suggesting that contaminating MEFs in the FACS experiment may account for much of this signature. However, a few candidate MEF genes were expressed at low levels throughout the iN cell populations (Extended Data Fig. 6c, d).

We also explored the diversity of iN cells by identifying genes that are differentially expressed between different iN cell populations using scRNA-seq. Among these were receptors, ion channels and transmembrane proteins, some of which have known roles in disease (Fig. 4c, Extended Data Fig. 6f). Individual iN cells within a population also exhibited mosaic expression of certain genes, either owing to inherent cellular diversity or high drop-out rates seen with scRNA-seq. However, *t*-distributed stochastic neighbour embedding (*t*-SNE) clustering of each individual transcription factor combination did not detect significant subclusters within populations, supporting a model in which iN cells exhibit low intra-population heterogeneity (Extended Data Fig. 6e). Therefore, while population RNA-seq data are sensitive to small subpopulations of contaminating cells, the overall predictions they make map well to the patterns of gene expression in individual iN cells. This suggests that each transcription factor pair induces a limited set of potentially related iN cell fates, as shown for induced sensory neurons[12].

Diversity among iN cells

To assess patterns of transcriptional diversity among the iN cell populations, we applied weighted gene co-expression network analysis (WGCNA) to generate co-expression modules[20,21]. As expected, several modules resembled the core transcriptome; they were enriched in most iN cells compared to in MEFs and included genes related to neuronal differentiation, synapses and metabolism (based on enriched GO terms)[22,23]. Other modules were expressed only in subsets of iN cells. Module 24 was enriched in iN cells generated with the POU4F (also known as Brn3) family of transcription factors, whereas two modules enriched for neurogenic genes were reciprocally expressed in iN cells generated with different bHLH family genes (Neurog versus Ascl families), with the latter including myogenic genes[19]. Other 'synergistic' modules were expressed in diverse iN cell populations generated with non-overlapping transcription factor pairs (Fig. 5a, Extended Data Fig. 7a, b, Supplementary Table 4). These analyses document multiple axes of diversity among iN cells and provide evidence for synergistic interactions between transcription factor pairs.

Producing iN cells with desired neurotransmitter expression profiles is of interest for translational medicine. Here we show that different iN cell populations express different neurotransmitter-related genes, such as those integral to excitatory, inhibitory and cholinergic identities. However, most of these genes are expressed at lower levels than in the endogenous neurons (but higher than in MEFs). This lower expression

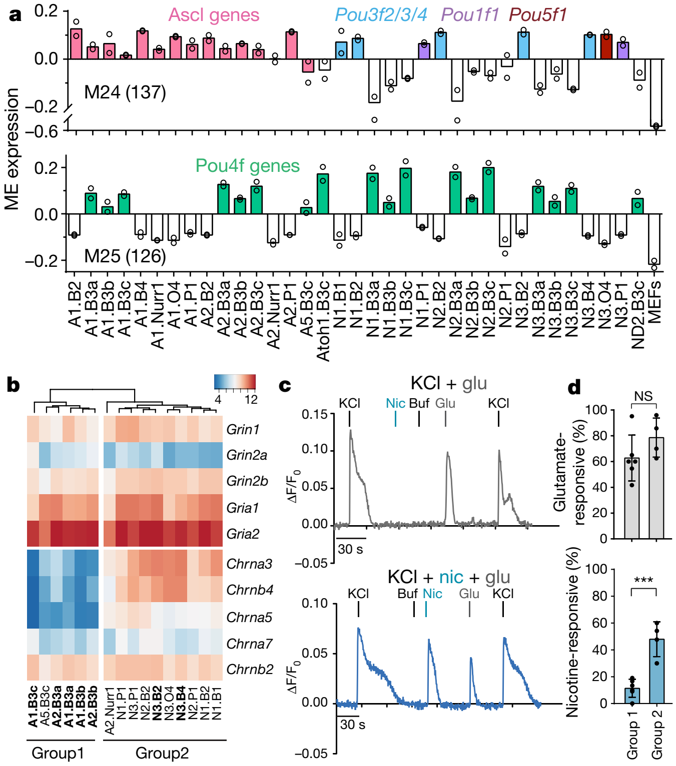


Fig. 5 | Transcriptional and functional diversity of iN cell populations. **a**, Average WGCNA module eigengene (ME) expression of the 35 iN cell populations (in duplicate) for representative modules correlated with POU subclasses (module 24 (M24), 137 genes assigned) and nonlinear or synergistic modules (module 25 (M25), 126 genes). Colours highlight populations generated with shared transcription factors. **b**, Expression of selected glutamate and nicotinic acetylcholine receptor subunit genes in iN cell populations, grouped using hierarchical clustering based on correlation distance. Expression is shown as the mean of two DESeq2 vsd-normalized RNA-seq counts. Bolded iN cell populations were used in calcium imaging experiments (Extended Data Fig. 8b). **c**, Calcium responses of two representative cells (group 1, top, grey; group 2, bottom, blue) to 100–250 mM KCl and 1 mM glutamate (glu), 100 μM nicotine (nic) and buffer alone (buf) plotted as $\Delta F/F_0$ versus time. **d**, Percentages of glutamate- and nicotine-responsive cells per group out of total KCl-responsive cells ($n = 218$ total cells). Group 1 and group 2 ($n = 6$ and 4 independendent experiments, respectively) comprise iN cell populations with the lowest and highest overall expression of nicotinic acetylcholine receptors, respectively. $^{***}P = 0.0004$, NS, not significant; unpaired Student's *t*-test. Data are mean ± s.d.

level could reflect mixed or immature iN cell populations; however, our detection of EPSCs, which involve presynaptic neurotransmitter release, indicates that these levels can have functional relevance (Extended Data Fig. 7c).

Previous studies showed that *Ascl1/Nr4a2* (alone or in combination with *Lmx1a*) can produce functional induced dopaminergic neurons[4,6,24–26]. Our results are partly consistent with these studies; we detected expression of tyrosine hydroxylase (*Th*) and dopamine decarboxylase (*Ddc*) genes and low levels of the noradrenergic or adrenergic genes (*Dbh, Slc6a2* (also known as *NET*)) in iN cells generated with *Ascl1* or *Ascl2* and *Nr4a2* compared to other *Th*-expressing iN cell populations. However, we did not reliably detect the dopamine transporter DAT (also known as Slc6a3) in any iN cell population, perhaps owing to a requirement for further maturation or a different combination of transcription factors (Extended Data Fig. 7d, e).

This reprogramming screen can be applied to generate iN cell populations that express neuroactive ligand receptors, ion channels or other genes related to human neurologic disease[27]. For example, human genetic variation at the *CHRNA3–CHRNB4–CHRNA5* nicotinic receptor gene cluster influences susceptibility to nicotine addiction, lung cancer and alcoholism[28]. We identified groups of iN cell populations that expressed similar levels of glutamate receptor subunits but differed in expression of *Chrna3*, *Chrnb4* and *Chrna5* (Fig. 5b, Extended Data Fig. 8a). Calcium imaging showed that while all iN cells responded equally to glutamate, iN cells enriched in *Chrna3, Chrnb4 and Chrna5* responded preferentially to nicotine, providing proof-of-principle that the iN cell dataset offers guidance for engineering iN cells that exhibit desired functional properties (Fig. 5c, d, Extended Data Fig. 8b).

Next, we compared the transcriptional profiles of iN cell populations to endogenous neural populations (Extended Data Figs. 9, 10). These analyses, including cell-type specific enrichment analysis (CSEA)[29,30], successfully distinguished between neuronal and glial identity and identified similarities between iN cells that resemble peripheral sensory neurons and endogenous DRG neurons. Several iN cell populations exhibited similarities to endogenous subpopulations, including habenula, hypothalamic and hippocampal neurons. However, in most cases, the sets of genes driving the relative similarity are relatively small and inconsistent with any conclusive assignment to a known endogenous subtype. This is perhaps unsurprising given the large number of neuronal subtypes, their inherent transcriptional heterogeneity, and the expected differences between neurons cultured in vitro versus those responding to the dynamic environment in a living animal.

Finally, we established a searchable database that links any gene to its pattern of expression among all iN cells and identifies other highly correlated genes (http://biogps.org/dataset/BDS_00016/). This 'look-up table' for producing iN cell populations expressing genes-of-interest can be used as a starting point for optimizing production of diverse iN cell subtypes or for pharmacogenomics, disease modelling and translational studies.

Discussion

Direct reprogramming has challenged the principles underlying plasticity of differentiated cell states established by decades of developmental studies by showing that transient expression of transcription factors can convert cell types across lineages. One interpretation of direct reprogramming studies is that only a small number of combinations of transcription factors may be capable of rewiring cell fate for each resulting lineage[31]. Our study suggests the opposite—more than 12% of reprogramming pairs we tested reprogrammed fibroblasts into neuron-like cells (iN cells).

Despite the use of many different combinations of transcription factors, and of transcription factors that are not recognized as promoting neuronal identity (such as POU5F1, also known as OCT4), all iN cells we tested exhibited key traits of differentiated neurons, including neuronal morphology, electrical excitability and expression of synaptic markers. The stabilization of these aspects of neuronal identity suggests the involvement of positive-feedback transcriptional networks that are actively maintained and arise in a cell-autonomous manner—independently of neural environment or developmental context. One interpretation of these data is that reprogramming may engage a restricted set of convergent transcriptional networks that mirror those used to establish or maintain aspects of neuronal diversity in vivo. Alternatively, iN cells may represent somewhat artificial cell types (as is the case for induced pluripotent stem cells); yet they may still be useful for interrogating the cellular function of neuronal genes in vitro. Though we favour the former hypothesis, additional transcriptional analyses of diverse sets of endogenous neurons and iN cells will be required to address this issue[32–34]. Finally, expanding the screen to include human transcription factors may enable production of novel human iN cell subtypes that are useful for deciphering human-specific aspects of neuronal cell biology and mechanisms related to human neurologic disease.

Online content

Any Methods, including any statements of data availability and Nature Research reporting summaries, along with any additional references and Source Data files, are available in the online version of the paper at https://doi.org/10.1038/s41586-018-0103-5.

Received: 18 May 2017; Accepted: 29 March 2018;

1. Davis, R. L., Weintraub, H. & Lassar, A. B. Expression of a single transfected cDNA converts fibroblasts to myoblasts. *Cell* **51**, 987–1000 (1987).
2. Takahashi, K. & Yamanaka, S. Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell* **126**, 663–676 (2006).
3. Vierbuchen, T. et al. Direct conversion of fibroblasts to functional neurons by defined factors. *Nature* **463**, 1035–1041 (2010).
4. Caiazzo, M. et al. Direct generation of functional dopaminergic neurons from mouse and human fibroblasts. *Nature* **476**, 224–227 (2011).
5. Chanda, S., Marro, S., Wernig, M. & Südhof, T. C. Neurons generated by direct conversion of fibroblasts reproduce synaptic phenotype caused by autism-associated neuroligin-3 mutation. *Proc. Natl Acad. Sci. USA* **110**, 16622–16627 (2013).
6. Kim, J. et al. Functional integration of dopaminergic neurons directly converted from mouse fibroblasts. *Cell Stem Cell* **9**, 413–419 (2011).
7. Pang, Z. P. et al. Induction of human neuronal cells by defined transcription factors. *Nature* **476**, 220–223 (2011).
8. Pfisterer, U. et al. Direct conversion of human fibroblasts to dopaminergic neurons. *Proc. Natl Acad. Sci. USA* **108**, 10343–10348 (2011).
9. Son, E. Y. et al. Conversion of mouse and human fibroblasts into functional spinal motor neurons. *Cell Stem Cell* **9**, 205–218 (2011).
10. Vadodaria, K. C. et al. Generation of functional human serotonergic neurons from fibroblasts. *Mol. Psychiatry* **21**, 49–61 (2016).
11. Wapinski, O. L. et al. Hierarchical mechanisms for direct reprogramming of fibroblasts to neurons. *Cell* **155**, 621–635 (2013).
12. Blanchard, J. W. et al. Selective conversion of fibroblasts into peripheral sensory neurons. *Nat. Neurosci.* **18**, 25–35 (2015).
13. Hockemeyer, D. et al. A drug-inducible system for direct reprogramming of human somatic cells to pluripotency. *Cell Stem Cell* **3**, 346–353 (2008).
14. Xu, C. et al. Immortalized fibroblast-like cells derived from human embryonic stem cells support undifferentiated cell growth. *Stem Cells* **22**, 972–980 (2004).
15. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
16. Dennis, G. Jr et al. DAVID: Database for Annotation, Visualization, and Integrated Discovery. *Genome Biol.* **4**, 3 (2003).
17. Schoenherr, C. J., Paquette, A. J. & Anderson, D. J. Identification of potential target genes for the neuron-restrictive silencer factor. *Proc. Natl Acad. Sci. USA* **93**, 9881–9886 (1996).
18. Drouin-Ouellet, J. et al. *REST* suppression mediates neural conversion of adult human fibroblasts via microRNA-dependent and -independent pathways. *EMBO Mol. Med.* **9**, 1117–1131 (2017).
19. Treutlein, B. et al. Dissecting direct reprogramming from fibroblast to neuron using single-cell RNA-seq. *Nature* **534**, 391–395 (2016).
20. Langfelder, P. & Horvath, S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* **9**, 559 (2008).
21. Zhang, B. & Horvath, S. A general framework for weighted gene co-expression network analysis. *Stat. Appl. Genet. Mol. Biol.* **4**, https://doi.org/10.2202/1544-6115.1128 (2005).
22. Mi, H., Poudel, S., Muruganujan, A., Casagrande, J. T. & Thomas, P. D. PANTHER version 10: expanded protein families and functions, and analysis tools. *Nucleic Acids Res.* **44** (D1), D336–D342 (2016).
23. Mi, H., Muruganujan, A., Casagrande, J. T. & Thomas, P. D. Large-scale gene function analysis with the PANTHER classification system. *Nat. Protoc.* **8**, 1551–1566 (2013).
24. Oh, S. I. et al. Efficient reprogramming of mouse fibroblasts to neuronal cells including dopaminergic neurons. *The Scientific World Journal* **2014**, 957548 (2014).
25. Addis, R. C. et al. Efficient conversion of astrocytes to functional midbrain dopaminergic neurons using a single polycistronic vector. *PLoS ONE* **6**, e28719 (2011).
26. Dell'Anno, M. T. et al. Remote control of induced dopaminergic neurons in parkinsonian rats. *J. Clin. Invest.* **124**, 3215–3229 (2014).
27. Albuquerque, E. X., Pereira, E. F., Alkondon, M. & Rogers, S. W. Mammalian nicotinic acetylcholine receptors: from structure to function. *Physiol. Rev.* **89**, 73–120 (2009).
28. Thorgeirsson, T. E. et al. A variant associated with nicotine dependence, lung cancer and peripheral arterial disease. *Nature* **452**, 638–642 (2008).
29. Dougherty, J. D., Schmidt, E. F., Nakajima, M. & Heintz, N. Analytical approaches to RNA profiling data for the identification of genes enriched in specific cells. *Nucleic Acids Res.* **38**, 4218–4230 (2010).
30. Xu, X., Wells, A. B., O'Brien, D. R., Nehorai, A. & Dougherty, J. D. Cell type-specific expression analysis to identify putative cellular mechanisms for neurogenetic disorders. *J. Neurosci.* **34**, 1420–1431 (2014).
31. Chin, M. T. Reprogramming cell fate: a changing story. *Front. Cell Dev. Biol.* **2**, 46 (2014).
32. Poulin, J. F., Tasic, B., Hjerling-Leffler, J., Trimarchi, J. M. & Awatramani, R. Disentangling neural cell diversity using single-cell transcriptomics. *Nat. Neurosci.* **19**, 1131–1141 (2016).
33. Usoskin, D. et al. Unbiased classification of sensory neuron types by large-scale single-cell RNA sequencing. *Nat. Neurosci.* **18**, 145–153 (2015).
34. Tasic, B. et al. Adult mouse cortical cell taxonomy revealed by single cell transcriptomics. *Nat. Neurosci.* **19**, 335–346 (2016).

Acknowledgements We thank M. Haynes, B. Seeger and A. Saluk for cell sorting, S. Head, J. Shimashita and J. Lesdesma for next-generation sequencing, K. Spencer for microscopy, V. Lo Sardo, W. Ferguson, M. Duran, J. Hazen, A. Adler and the Topol laboratory for technical assistance, R. Vega Perez for cell counting, and A. Su and J. Fouquier for assistance with BioGPS. BioGPS work is funded by R01 GM083924 to A. Su. This research was supported the National Brain Research Program of Hungary (KTIA_NAP_13-2014-0018 to A.S.), by the NIH (NIDA, DA031566 to P.P.S.), by The Scripps Translational Science (A.T.), (CTSA; 5 UL1 TR001114 to A.T.), (U54GM114833 to A.T.), (NIDCD, DC012592 to K.K.B.), (NIMH, MH102698 to K.K.B.), (NIA, DP1 AG055944), and the Dorris Neuroscience Center (K.K.B.), a pre-doctoral fellowship from CIRM (J.W.B., R.T. and S.L.), an NSF Predoctoral Fellowship (R.T.) and the Andrea Elizabeth Vogt Memorial Award (J.W.B.).

Author contributions K.K.B., R.T. and S.L. designed and conceived the experiments, wrote and revised the manuscript and all of the authors edited the final drafts. K.T.E., J.W.B. and R.T. performed and analysed the screen experiments. A.S., I.S. and P.P.S. performed electrophysiology. R.T., S.L. and P.C. prepared cDNA libraries. R.T., S.L., P.C. and A.T. performed RNA-seq analysis. J.B. performed HOMER motif enrichment analysis. C.W. made data available through BioGPS.org. R.T. and S.L. performed all remaining experiments.

Competing interests The authors declare no competing interests.

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0103-5.

Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0103-5.

Reprints and permissions information is available at http://www.nature.com/reprints.

Correspondence and requests for materials should be addressed to K.K.B.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

The Institutional Animal Care and Use Committees of The Scripps Research Institute approved all animal procedures. As experimental comparison of groups of animals was not conducted in this study, sample size choice, randomization, sex of animals and blinding were not relevant for animal procedures. Human iN cells were produced from induced pluripotent stem cell lines generated from de-identified donors; the cells were validated in the Baldwin laboratory and are routinely tested for mycoplasma. Investigators were blinded to perform the induced neuronal cell counts in the original screen which were replicated in some cases by non-blinded investigators. In no other experiments were investigators blinded.

Embryonic fibroblast isolation and derivation. Wild-type CD1 mice and heterozygous tau–eGFP mice (Jackson Laboratory, *Mapt*$^{tm1(EGFP)Klt}$/J, 004779) were bred at The Scripps Research Institute animal facility. Mouse embryonic fibroblasts (MEFs) were isolated under a dissection microscope from E13.5 embryos by removing the heads, limbs, internal organs and spinal columns to eliminate neurogenic cells. The remaining tissue was manually dissociated with 0.25% trypsin (Gibco) for 20 min at 37 °C. The trypsin was subsequently diluted with MEF medium (DMEM, 10% FBS and penicillin–streptomycin) and removed via centrifugation. Pelleted cells were re-suspended in MEF medium and seeded on gelatin-coated (0.01%) tissue culture plates. MEFs were grown to confluence and passaged at least twice before use.

Primary TTFs were isolated from 2–4-mm-long tail tips of P3 mouse pups. Tail tips were first rinsed in 70% ethanol, washed with HBSS (Invitrogen), chopped into smaller pieces and dissociated 0.25% trypsin for 60 min at 37 °C. Subsequent steps are the same as in the MEF isolation protocol.

For derivation of human embryonic fibroblasts (HEFs), human iPSCs colonies were collected using 0.5 mM EDTA (Invitrogen) and differentiated by embryoid body formation. For human iPSCs, normal blood donor (NBD) lines previous generated in an independent study[35] were used and approved by the Scripps Institutional Review Board (IRB-11-5676). The embryoid bodies were cultured for seven days in non-adherent suspension culture dishes (Corning), two days in mTeSR medium (StemCell Technologies) and the following five days in DMEM containing 10% FBS (vol/vol). On day 8, the embryoid bodies were plated onto adherent tissue culture dishes and passaged according to primary fibroblast protocols using 0.25% trypsin for two to three passages before the start of experiments.

Molecular cloning, cell culture and lentiviral transduction. cDNAs for transcription factors were cloned into lentiviral constructs under the control of the tetracycline operator (TetO). The cDNA for *POU4F1* (previously known as *BRN3A*) and *NHLH2* (previously *HEN2*) are the only human transcription factors we used; the human BRN3A protein shares 97% homology with the mouse protein. The *BRN3A* gene was cloned as described[12]. Replication-incompetent VSVg-coated lentiviral particles were packaged in 293 T cells (ATCC), collected 48 h after transfection, and filtered through a 45-μm PVDF membrane before use. HEK 293 T cells were tested and authenticated by ATCC before experiments. The reprogramming method is a modification of a previously described protocol[12]. Passage two MEFs were infected with lentivirus in MEF medium. After 12–24 h of infection, virus-containing medium was replaced with fresh MEF medium. Transcription factors were induced 48 h after infection by switching to MEF medium supplemented with 5 μM doxycycline (Sigma). Four days after initiating induction with doxycycline, MEF medium was replaced with N3 medium[3], but using N2 supplement (Gibco) in place of some components. Eight days after induction, doxycycline was withdrawn. Ten days after induction, the cells were switched to neural maintenance medium, which consisted of a 1:1 mix of N3 medium and Neurobasal medium (Invitrogen) supplemented with B27 (minus vitamin A, Gibco) and bFGF (10 ng/ml) (N3/NB medium). Efficiency of conversion was measured by the number of TUJ1$^+$ cells divided by the initial number of plated cells. Transcription factor pairs that included bHLH factors *Ascl1*, *Ascl2*, *Neurog1* and *Neurog3* were normalized by subtracting the percentage of TUJ1$^+$ cells generated from the bHLH factors alone (range 0.01–0.39%).

Immunohistochemistry. Cells for immunofluorescence staining were fixed with 4% paraformaldehyde for 10 min at room temperature. Cells were then washed three times with phosphate-buffered saline and subsequently blocked in 5% horse serum and 0.1% Triton X-100 (Sigma) for 1 h at room temperature. Primary staining was performed overnight at 4 °C in the blocking buffer. Cells were again washed three times and then stained with secondary antibodies diluted in blocking buffer for 1 h at room temperature. The following primary antibodies and dilutions were used: TUJ1 (Sigma-Aldrich T2200, Rabbit, polyclonal, 1:500), MAP2 (Sigma-Aldrich M4403, mouse, HM-2, monoclonal, 1:500), and synapsin 1 (Synaptic Systems 106103, Rabbit, polyclonal, 1:500). Antibodies were validated using mouse primary neurons and tissue as a positive control and mouse embryonic fibroblasts as a negative control.

Electrophysiology. MEFs from tau–eGFP mice were reprogrammed and cultured as described on Thermanox plastic coverslips (33 mm diameter). Coverslips were placed in the recording chamber mounted on an Olympus BX51 microscope. To identify tau–eGFP$^+$ cells that expressed synapsin, we transduced candidate iN cells with lentivirus encoding the fluorescent red protein TdTomato, under the control of a *Syn1* promoter. Spontaneous activity and evoked responses were recorded from identified cells at day 16 to 24 post-induction under whole-cell patch clamp at 33 °C. Similar to the electrophysiology protocol described[12], signals were amplified using a MultiClamp700B (Molecular Devices) and acquired using the data acquisition software DASYLab v.11 (National Instruments) at 20 kHz. Patch pipettes with input resistances of 6–8 MΩ were pulled from standard wall glass of 1.5-mm OD (Warner Instruments) and filled with solution containing 120 mM potassium gluconate, 10 mM KCl, 10 mM HEPES, 10 mM EGTA, 2 mM MgATP, 0.3 mM Na_3GTP at pH 7.3. The bath solution (artificial cerebrospinal fluid) was composed of 125 mM NaCl, 2.5 mM KCl, 2 mM $CaCl_2$, 1 mM $MgCl_2$, 1.25 mM NaH_2PO_4, 26 mM $NaHCO_3$ and 25 mM glucose. To record voltage responses of the identified iN cells, we used incremental levels of constant, rectangular current steps of 350-ms duration. The initial current step level was −50 to −200 pA, depending on the observed input resistance of the cell. Steps were incremented by + 2 or + 5 pA in successive cycles of stimulation at a rate of 1 Hz. Analysis of the evoked responses was performed in software developed by A. Szücs (IVAnalyzer). For each cell, several physiological parameters, including the resting membrane potential, rheobase, input resistance at rest and spike amplitude, were measured.

Spontaneous postsynaptic potentials were occasionally observed in the recorded iN cells. We performed voltage-clamp recordings of postsynaptic current whenever such activity was detected (150–200-s recordings at −50 mV holding potential). At this potential, the inward currents we observed were identified as EPSCs, considering the typical resting membrane potential of the iN cells (near −50 mV). GABAergic inputs do not typically produce such prominent EPSCs at this holding potential.

Human iN cells generated from HEFs were also identified using the *SYN1*-TdTomato reporter virus. Recordings were performed between 26 and 31 days after induction with doxycycline. Voltage-gated currents were induced by 400-ms voltage steps to −115 mV to −5 mV in 10-mV increments from the initial potential of −65 mV. The leak currents were subtracted from the voltage-gated currents before analysis. Leak currents were calculated using the currents induced by stimulation from −65 mV to −55 mV and scaling them to the corresponding membrane voltage. Whole-cell currents were filtered at 2 kHz and sampled at 20 kHz with a Digidata 1440 interface controlled by pClamp Software (Molecular Devices).

Cell sorting. Reprogrammed candidate iN cells generated from heterozygous tau–eGFP MEFs were prepared for FACS by first detaching cells from the culture plate using Accutase (Innovative Cell Technologies). Accutase was subsequently diluted with neural maintenance media (N3/NB media) and removed by centrifugation. Pelleted cells were resuspended in neural maintenance medium, triturated and strained through 35-μm nylon mesh filter to obtain single-cell suspensions. Viability markers DAPI (1 μM) and DRAQ5 (1 μM, BioStatus DR50050) were added to the suspension at least 10 min before sorting. Appropriate gates for FACS were set based on tau–eGFP, DAPI and DRAQ5 intensities to isolate live tau–eGFP$^+$ cells as shown in Extended Data Fig. 4c–d using the MoFlo Astrios (Beckman Coulter). Isolated cells were sorted into TRIzol LS (Invitrogen).

Similarly, endogenous neuronal populations were isolated from the appropriate transgenic reporter mice at postnatal day 21 (Extended Data Fig. 4i–o, Supplementary Table 2). Dissected tissue samples were dissociated as described[36], with the following modifications. Manual homogenization was conducted with a scalpel rather than with a tissue slicer. We also used papain-containing L-cysteine (PAP2 10 U ml^{-1}, Worthington Biochemical) because its higher activity allowed for shorter dissociation times[37] (15 min total). During papain digestion, samples were triturated every 5 min using P1000 plastic tips instead of siliconized Pasteur glass pipettes. After centrifugation using the density gradient, we found viable neurons in the fraction containing the cell pellet and the fraction 2 ml immediately above the pellet. Both fractions were combined and washed once in 10 ml HAGB (Hibernate-A (Gibco A1247501), 1 × B-27 supplement (Gibco 12587010), 500 μM GlutaMAX (Gibco 35050061)). After subsequent centrifugation, pelleted cells were resuspended in HAGB, filtered and kept on ice until FACS sorting. As with the candidate iN cells, viability markers DAPI and DRAQ5 were added to the suspension and appropriate gates were set to purify cells into TRIzol LS.

Population RNA isolation. Total RNA was isolated from FACS-sorted cells using Direct-zol RNA MiniPrep Kit (Zymo Rsearch) according to the manufacturer's protocol, except linearized acrylamide (1 μg) was added to each sample before the first step and Zymo-Spin IC columns were used in replacement of Zymo-Spin IIC columns. RNA quality and quantity was determined with an Agilent 2100 Bioanalyzer. RNA integrity numbers (RINs) for all iN cell samples were between 6 and 10 (median = 8.7). The amount of RNA per sorted event was between 1 and 15 pg (median = 7.9 pg). Therefore, approximately 1,500 to 2,000 cells were required to yield 10 ng RNA for library input.

Population RNA-seq library preparation and sequencing. RNA-seq libraries were prepared from duplicate populations of tau–eGFP$^+$ cells generated from

35 different transcription factor pairs 16 days after induction. Additional sequenced populations included biological replicates of MEFs, representative endogenous postnatal-day-21 neural populations, and whole-brain samples. Typically 10 ng purified, high quality RNA served as input for SMARTer Ultra Low Input RNA Kit for Sequencing v3 (Clontech Laboratories). A few replicate libraries were prepped from 1–7 ng of input total RNA (Supplementary Table 2). These were comparable to libraries prepped from 10 ng RNA since correlation coefficients were greater than 0.98 between libraries prepped from 1, 5 and 10 ng of the same total RNA (Extended Data Fig. 2f–h). Amplified cDNA was assessed for quality using High Sensitivity DNA Kit (Agilent Technologies) and sheared using the Covaris system. Sequencing libraries were subsequently prepped using NEBNext Ultra DNA Library Prep Kit for Illumina. Seventy-five-base-pair single-end reads generated using Illumina's NextSeq platform were mapped to the mouse genome (UCSC mm10 database) by first removing adapters and low quality bases using Trimmomatic (v0.32, ILLUMINACLIP: TruSeq3-SE.fa:2:30:10 LEADING:3 TRAILING:3)[38]. Reads were then aligned using STAR[39] and counts were generated using HTSeq[40]. Mm10 did not include *Ascl5*; therefore, we added it to the reference GTF file in HTSeq. It is also important to note that some libraries were prepared using SMARTer Ultra Low Input RNA for Illumina Sequencing - HV (Clontech Laboratories) and sequenced on Illumina's HiSeq platform, resulting in 100-bp single reads. Libraries were sequenced to a mean of ~37.5 million uniquely mapped 75-bp single-end reads per replicate (Supplementary Table 2).

RNA-seq data analysis (DESeq2 and principal component analysis). RNA-seq data were analysed using several R[41] packages available through Bioconductor[42]. Differential gene expression analysis was conducted using DESeq2[15]. Heat maps were generated using gplots[43]. PCA plots were drawn using rgl[44] and pca3d[45].

Ingenuity upstream regulator analysis. Ingenuity upstream regulator analysis in Qiagen's Ingenuity Pathway Analysis (IPA) was used to identify the cascade of upstream regulators of the core gene set. IPA utilizes a priori knowledge of expected interactions between transcriptional regulators and their target genes stored in Ingenuity Knowledge Base, a scientific literature-based database.

Hypergeometric optimization of motif enrichment analysis. To determine the regulatory elements acting within the iN cells, specifically transcription factors, motif enrichment was performed on the promoters of differentially expressed genes in the iN cells. The known motif enrichment routine in the findMotifs.pl routine available in HOMER software was used to perform the analysis[46]. Known motif enrichment in HOMER is performed by scanning a defined set of promoter regions for motifs defined by a set of position weight matrices (PWMs) and using ZOOPS (zero or one occurrence per sequence) counting coupled with a hypergeometric enrichment test to determine significance. Built into HOMER is a curated set of binding site motifs taken from the TRANSFAC database[47]. To expand our search entries in JASPAR core[48], a curated collection of transcription factor binding profiles was converted into PWMs for use in the analysis. HOMER asks for a threshold to be set for all PWMs. This threshold determines the minimum log odds score that is allowed for a sequence to be considered a match with the motif described in a given PWM. When converting the JASPAR profiles, the threshold was set by allowing for the least likely base in the most likely mismatched nucleotide of each motif, which was chosen because it allowed for some degeneracy when searching for possible transcription factor binding sites while excluding overly mismatched sequences.

Several promoter sets are available within HOMER. We used the mm9 genome build with a promoter region defined as 2,000-bp upstream and 50-bp downstream of the transcription start site for all identified genes in the mm9 build. The background gene set was restricted to those genes that were detectable in the RNA-seq experiments, excluding those genes whose transcripts had fewer than 100 reads mapped to them across all datasets. All other parameters available for findMotifs.pl were left as their defaults.

Single-cell RNA-seq. For scRNA-seq, iN cells from each transcription factor combination were sorted by FACS as described above and loaded as independent suspension samples on a GemCode Single-Cell Instrument (10 × Genomics) which utilizes a droplet-based method. Single-cell libraries were generated using the commercially available Chromium Single Cell 3′ V2 Reagent Kit (10 × Genomics; PN-120237, PN-120236, PN-120262) following the manufacturer's protocol[49]. Amplified cDNA and subsequent libraries were assessed for quantity and quality using High Sensitivity DNA Kit on an Agilent 2100 Bioanalyzer (Agilent Technologies). Libraries were sequenced as 150-bp paired-end reads on Illumina's NextSeq platform. Samples were de-multiplexed, aligned and analysed using the 10 × Genomic CellRanger pipeline with default parameters. This pipeline implements STAR as an alignment tool. Multiple libraries were aggregated using the 'cellranger aggr–normalized = mapped (default)' command which sub-samples mapped reads so that each iN cell combination would have the same effective sequencing depth. Post-normalization mean reads per cell was 428,208. Visualization of *t*-SNE plots and UMI counts, and gene normalization and filtering of unexpressed genes were performed using the R package 'cellrangerRkit'. Violin plots for individual genes were made using the R package 'ggplot2' from log-transformed gene-barcode matrices.

Cell type-specific expression analysis. Cell type-specific expression analysis (CSEA) was done using the publicly available CSEA web-based tool provided[30] by the Dougherty laboratory (http://genetics.wustl.edu/jdlab/csea-tool-2/, Version 1.0: updated 11th October 2013). Uniquely enriched genes of individual iN cell populations served as the input candidate gene lists. Uniquely enriched genes were defined as genes that were significantly enriched (P-adjusted value < 0.05) in each iN cell population versus all other iN cell populations and MEFs as determined by DESeq2 (Supplementary Table 4). Overlaps of these gene lists with a particular cell type or region for which data are currently available were identified by Fisher's exact test with Benjamini–Hochberg correction.

Weighted gene co-expression network analysis. Weighted gene co-expression network analysis (WGCNA) has previously been described in detail[21] and summarized in papers using this technique[50]. DESeq2 vsd-normalized counts of all iN cell and MEF population replicates ($n = 72$) served as input into a user-friendly WGCNA R library[20]. To reduce the noise from low-expressing genes in our dataset, we only included genes in which the non-normalized counts were greater than 200 in at least one iN cell or MEF population, in both replicates ($n = 12{,}549$). We constructed a signed network, with a power of 12, using the default parameters except deepSplit = 4 and cutHeight = 0.999. Modules were merged if their module eigengenes were correlated with $R > 0.8$. Module hub genes were those that had the highest module membership (k_{ME}) for that module, which was calculated as the Pearson correlation between the gene and the corresponding ME.

Calcium imaging. Calcium imaging was performed 16–24 days after induction on iN cells transduced with a MAP2::GCaMP5G lentiviral reporter[25] and a *Syn1*-TdTomato lentiviral reporter. Imaging was performed in Tyrode's solution (145 mM NaCl, 2.5 mM KCl, 10 mM Hepes, NaH_2PO4, 2 mM CaCl2, 1 mM MgCl2, 10 mM Glucose and 0.4 mM ascorbic acid) at a constant flow rate. In a randomized order, we serially exposed the iN cells to 1 mM glutamate and 100 μM nicotine by direct application to the area of interest. We only analysed *Syn1*–TdTomato$^+$ cells that responded to transient exposure to 100–250 mM KCl at the beginning and end of each recording to ensure iN cells exhibited neuronal identity and maintained functional viability throughout the recording. Additionally, we did not include mechanosensitive cells that responded to Tyrode's solution alone. Similar to as previously described[12], calcium responses were calculated as the change in fluorescence intensity (ΔF) over the initial fluorescence intensity $(F - F_0)/F_0$, in which F is the fluorescence at a given time point and F_0 was calculated as the average of the first five unstimulated fluorescence measurements at the start of imaging. A non-response area for each recording was measured for background subtraction. The threshold for a positive calcium response to the addition of a ligand was determined as one $(F - F_0)/F_0$ greater than 0.01 in a 10-s window.

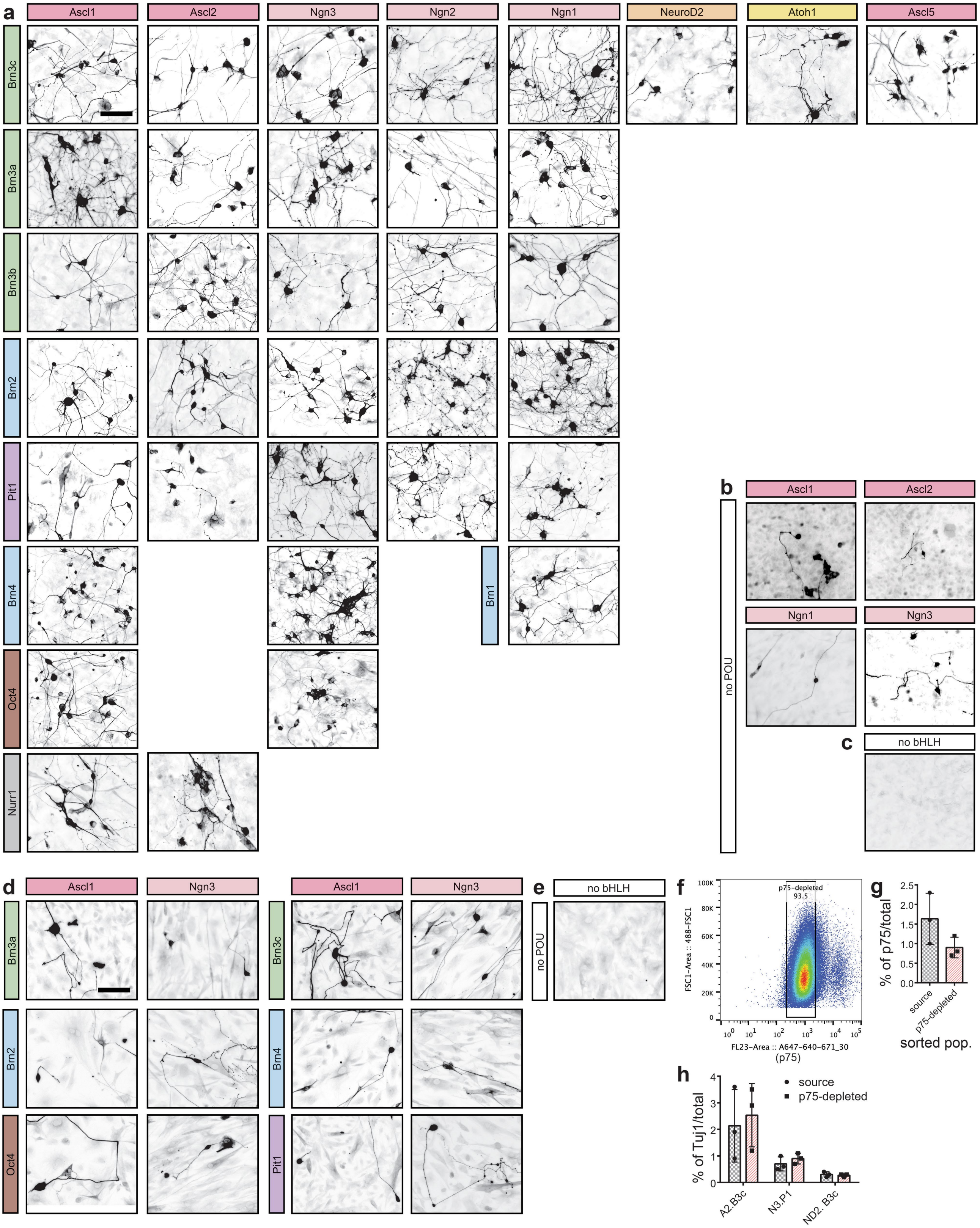
Statistics. Statistical analyses were performed using GraphPad Prism7 and detailed in the corresponding figure legends. Data from electrophysiology experiments were analysed by one-way ANOVA followed by Bonferroni's multiple comparison post hoc test. Similarity of variance between groups was confirmed by Brown–Forsythe test. Data from calcium imaging experiments were analysed by unpaired Student's *t*-test. Similarity of variance between groups was confirmed by *F* test.

Reporting summary. Further information on experimental design is available in the Nature Research Reporting Summary linked to this paper.

Data availability. RNA-seq data that support the findings of this study have been deposited to GEO with the accession code: GSE112381 and the BioGPS platform (http://biogps.org/dataset/BDS_00016/).

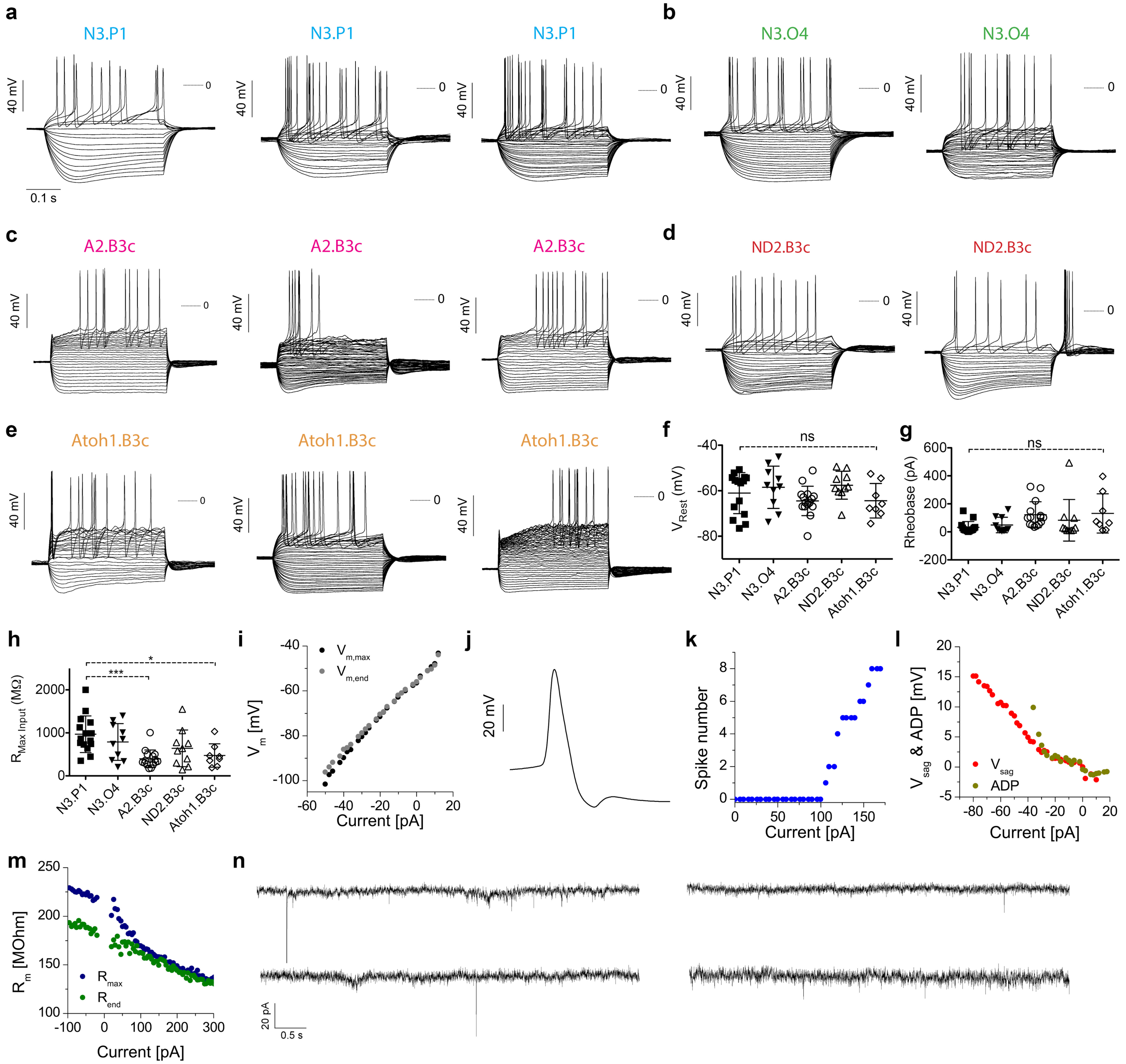
35. Lo Sardo, V. et al. Influence of donor age on induced pluripotent stem cells. *Nat. Biotechnol.* **35**, 69–74 (2017).
36. Brewer, G. J. & Torricelli, J. R. Isolation and culture of adult neurons and neurospheres. *Nat. Protoc.* **2**, 1490–1498 (2007).
37. Hazen, J. L. et al. The complete genome sequences, unique mutational spectra, and developmental potency of adult neurons revealed by cloning. *Neuron* **89**, 1223–1236 (2016).
38. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
39. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
40. Anders, S., Pyl, P. T. & Huber, W. HTSeq—a Python framework to work with high-throughput sequencing data. *Bioinformatics* **31**, 166–169 (2015).
41. R Core Team. *R: A Language and Environment for Statistical Computing* http://www.R-project.org/ (R Foundation for Statistical Computing, Vienna, Austria, 2013.
42. Gentleman, R. C. et al. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol.* **5**, R80 (2004).
43. Warnes, G. R. gplots: Various R Programming Tools for Plotting Data. https://rdrr.io/cran/gplots/ (2015).
44. Adler, D. et al. rgl: 3D Visualization Using OpenGL. https://rdrr.io/rforge/rgl/ (2016).

45. pca3d: Three Dimensional PCA Plots. https://cran.r-project.org/web/packages/pca3d/index.html (2015).
46. Heinz, S. et al. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol. Cell* **38**, 576–589 (2010).
47. Matys, V. et al. TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. *Nucleic Acids Res.* **34**, D108–D110 (2006).
48. Mathelier, A. et al. JASPAR 2016: a major expansion and update of the open-access database of transcription factor binding profiles. *Nucleic Acids Res.* **44** (D1), D110–D115 (2016).
49. Zheng, G. X. et al. Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.* **8**, 14049 (2017).
50. Hawrylycz, M. et al. Canonical genetic signatures of the adult human brain. *Nat. Neurosci.* **18**, 1832–1844 (2015).

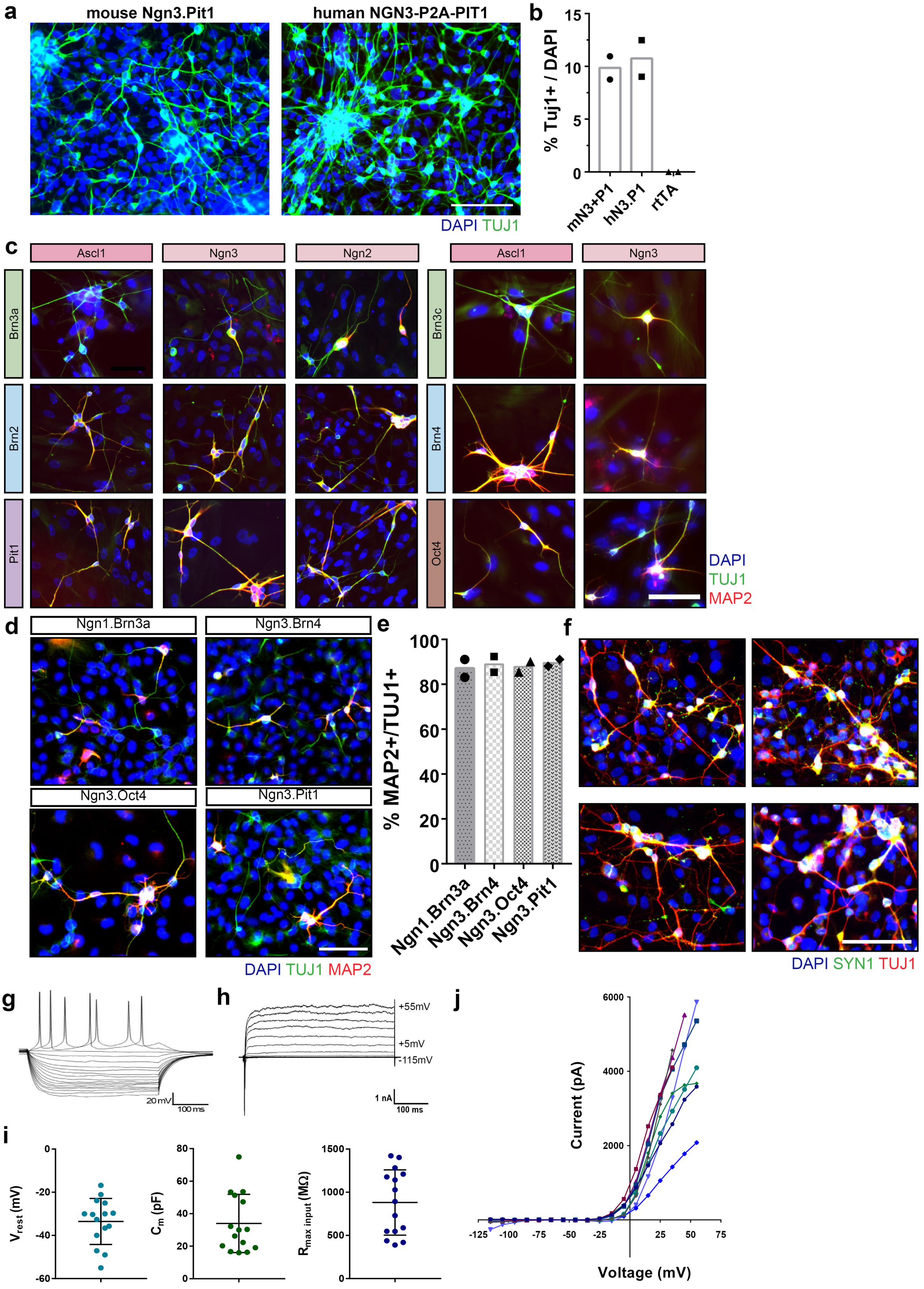


Extended Data Fig. 1 | See next page for caption.

Extended Data Fig. 1 | TUJ1 immunostaining of MEF- and TTF-derived iN cells and the p75-depletion experiment. **a**, TUJ1 immunofluorescence labelling on day 14–16 post-induction of 35 of the 76 hits that were selected for whole-transcriptome analysis. $n = 3$ independent experiments. **b**, TUJ1 immunofluorescence labelling of conditions with individual bHLH factors *Ascl1*, *Ascl2*, *Neurog1* and *Neurog3*. $n = 3$ independent experiments. **c**, TUJ1 immunofluorescence labelling of MEFs treated with only reverse tetracycline-controlled transactivator (rtTA), without reprogramming factors. $n = 3$ independent experiments. **d**, TUJ1 immunofluorescence labelling of TTFs derived from three-day-old mice and transduced with selected reprogramming combinations following the same reprogramming methods used with MEFs. Fixed and stained on day 16 post-induction. $n = 1$ independent experiment. **e**, TUJ1 immunofluorescence of TTFs treated with only rtTA, without reprogramming factors, and fixed and stained on day 16 post-induction. $n = 1$ independent experiment. **f**, Representative FACS gates of MEFs (~180,000 cells shown). MEFs were depleted of p75$^+$ neural crest cells by first gating for DAPI$^-$ cells (not shown) and collecting only those that were p75$^-$ (~93% of the DAPI$^-$ population). **g**, Quantification of immunostaining for p75$^+$ cells in source and p75-depleted MEF populations after expansion for four days after FACS, on the day of transduction for reprogramming. Data are mean ± s.d., $n = 3$ biologically independent samples. **h**, Percentage of TUJ1$^+$ cells derived from source and p75-depleted MEF populations 16 days after induction. A2, *Ascl2*; N3, *Neurog3*; ND2, *NeuroD2*; B3c, *Pou4f3*; P1, *Pou1f1*. Data are presented as the mean ± s.d., $n = 3$ biologically independent samples. Percentages of TUJ1$^+$ cells were not significantly different between source and p75-depleted conditions (two-way ANOVA, Sidak's multiple comparison test. A2.B3c, $P = 0.895$; N3.P1, $P = 0.985$; ND2.B3c, $P > 0.999$). Scale bars, 100 μm.

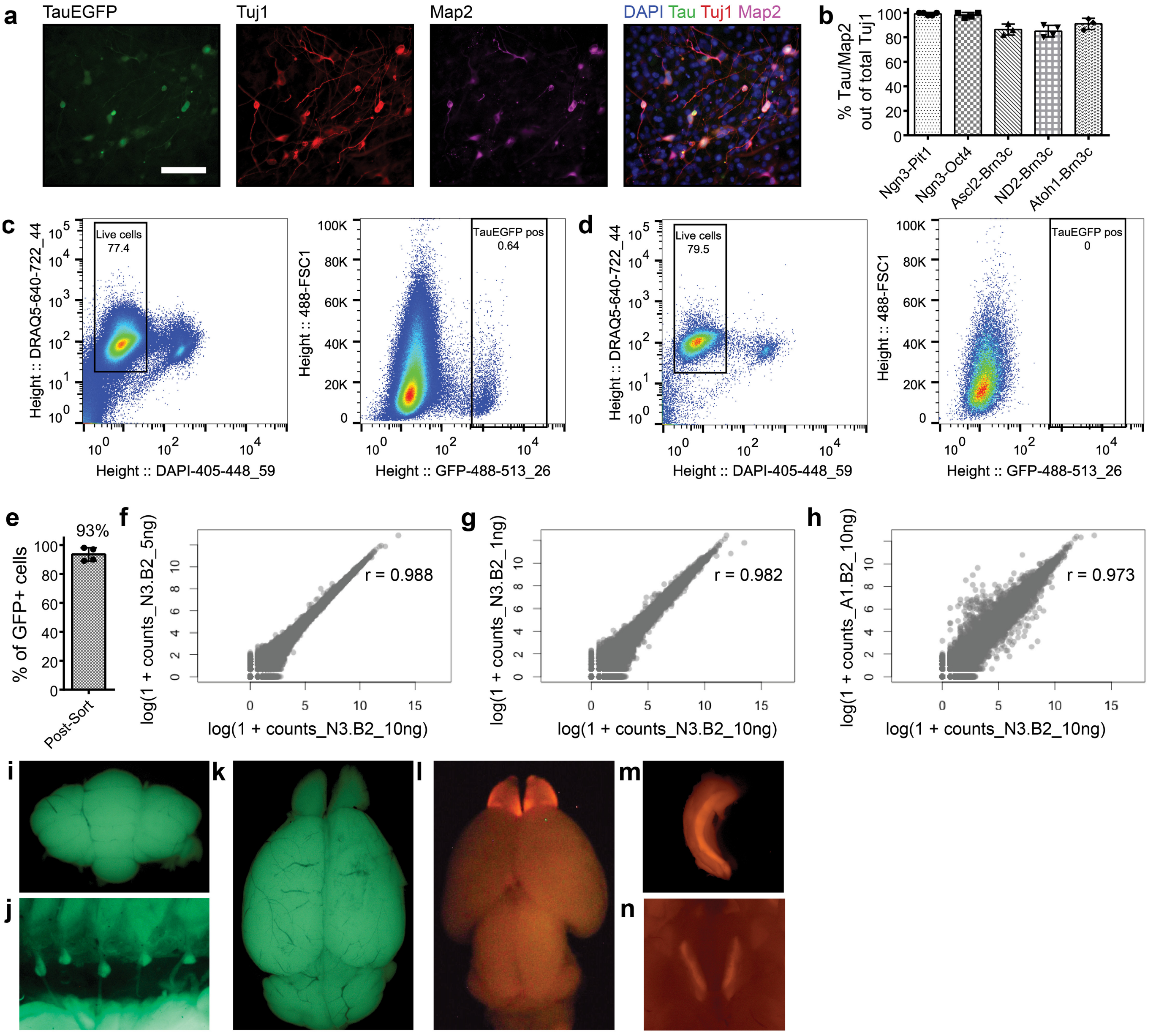


Extended Data Fig. 2 | Additional electrophysiological recordings of iN cells from five transcription factor combinations. **a**–**e**, Example voltage responses of representative iN cells from five transcription factor combinations: *Neurog3/Pou1f1* (**a**; $n = 3$ cells), *Neurog3/Pou5f1* (**b**; $n = 2$ cells), *Ascl2/Pou4f3* (**c**; $n = 3$ cells), *Neurod2/Pou4f3* (**d**; $n = 2$ cells) and *Atoh1/Pou4f3* (**e**; $n = 3$ cells). Cells were stimulated using incremental levels of intracellular current starting at −100 to −50 pA and reaching levels where intense firing of action potentials was observed. **f**–**g**, Quantification of resting membrane potential (**f**), rheobase (**g**) and membrane input resistance (**h**) for cells that exhibited current-induced action potentials. *Neurog3/Pou1f1* (N3.P1, $n = 15$ cells), *Neurog3/Pou5f1* (N3.O4; $n = 10$ cells), *Ascl2/Pou4f3* (A2.B3c; $n = 15$ cells), *Neurod2/Pou4f3* (ND2.B3c; $n = 10$ cells) and *Atoh1/Pou4f3* (Atoh1.B3c; $n = 8$ cells). Data are mean ± s.d.; ***$P = 0.0006$, *$P = 0.0228$; ns, not significant. One-way ANOVA, Tukey's multiple comparison test. **i**–**m**, Physiological properties of the cells. **i**, Current–voltage relationship obtained by plotting the observed membrane potential as a function of the injected current of both maximal voltage deflections (black) and the membrane potential at the end of the current step (grey). Data from the third *Neurog3/Pou1f1* cell in **a**. **j**, Selected action potential of the second *Neurog3/Pou5f1* cell in **b**. The dual spike after-hyperpolarization is indicative of Ca-dependent K^+ currents in this neuron. **k**, Input−output curve of the number of spikes as a function of the injected current. This cell starts firing at + 100 pA (rheobase). **l**, Plot of the voltage sag (red) and after depolarization (dark yellow) as a function of the current. The *Neurod2/Pou4f3* cells in **d** exhibit characteristic voltage sags under negative currents. The second *Neurod2/Pou4f3* cell also produces post-inhibitory rebound spikes. **m**, Plot of membrane resistance versus current. Blue symbols are resistance values calculated from maximal voltage deflections and green symbols were obtained from voltage levels just before the termination of the current step of the third *Atoh1/Pou4f3* cell in **e**. The decrease of membrane resistance as a function of current indicates the action of potent outward-rectifying K^+ currents. **n**, Representative current traces from four cells showing EPSCs from tau–eGFP$^+$, synapsin$^+$ cells generated with *Neurog3/Pou5f1* and *Neurod2/Pou4f3*.



Extended Data Fig. 3 | See next page for caption.

Extended Data Fig. 3 | Transcription factor pairs generate functional human iN cells from HEFs. **a**, Representative TUJ1 immunofluorescence labelling of human iN cells reprogrammed from HEFs using mouse *Neurog3/Pou1f1* or human *NEUROG3/POU1F1*. Scale bar, 100 μm. **b**, Quantification of TUJ1$^+$ DAPI$^+$ cells for mouse and human iN cells derived from mouse (m) or human (h) *Neurog3* and *Pou1f1* or rtTA only. Data from $n = 2$ biologically independent samples. **c**, Representative images of human iN cells reprogrammed from HEFs using pairs of mouse transcription factors. TUJ1 and MAP2 immunofluorescence labelling of 15 of the 76 positive pairwise combinations derived from the unbiased mouse screen. Fixed and stained on day 16–18 post-induction. Scale bar, 100 μm. Repeated with $n = 2$ independent experiments. **d**, Representative images of human iN cells reprogrammed from HEFs in an independent experiment from **c**. TUJ1 and MAP2 immunofluorescence labelling of four pairwise mouse transcription factor combinations. Fixed and stained on day 18 post-induction. Scale bar, 100 μm. **e**, Percentage fraction of MAP2$^+$TUJ1$^+$ cells from the four transcription factor combinations represented in **d**. Imaging from $n = 2$ biologically independent samples, 100 fields of view each. Number of TUJ1$^+$ cells is as follows: *Neurog1/Pou4f1* ($n = 166$ cells); *Neurog3/Pou3f4* ($n = 343$ cells); *Neurog3/Pou1f1* ($n = 235$ cells); *Neurog3/Pou5f1* ($n = 146$ cells). Data are mean ± s.d. **f**, Representative synapsin (SYN1) and TUJ1 immunofluorescence labelling of human iN cells reprogrammed with *Neurog3/Pou1f1* (91.5% positive for both). Scale bar, 100 μm. Repeated with $n = 3$ biologically independent samples. **g**–**j**, Electrophysiological recordings were performed on human iN cells generated with mouse *Neurog3/Pou1f1* between 26 and 31 days post-induction. **g**, Representative voltage responses from a *Syn1*–TdTomato$^+$ cell with neuronal morphology; 21 of 27 fluorescent cells tested (77%) generated action potentials upon current injection. **h**, Representative whole-cell currents evoked by hyperpolarizing and depolarizing voltage steps delivered from a holding potential of −65 mV. **i**, Passive membrane properties of human iN cells. Quantification of resting membrane potential (left), capacitance (middle) and membrane resistance (right) is shown as mean ± s.d. ($n = 15$ cells). **j**, Steady-state currents versus voltage in individual cells reflect the expression of depolarization-induced voltage-gated outward currents ($n = 9$ cells).

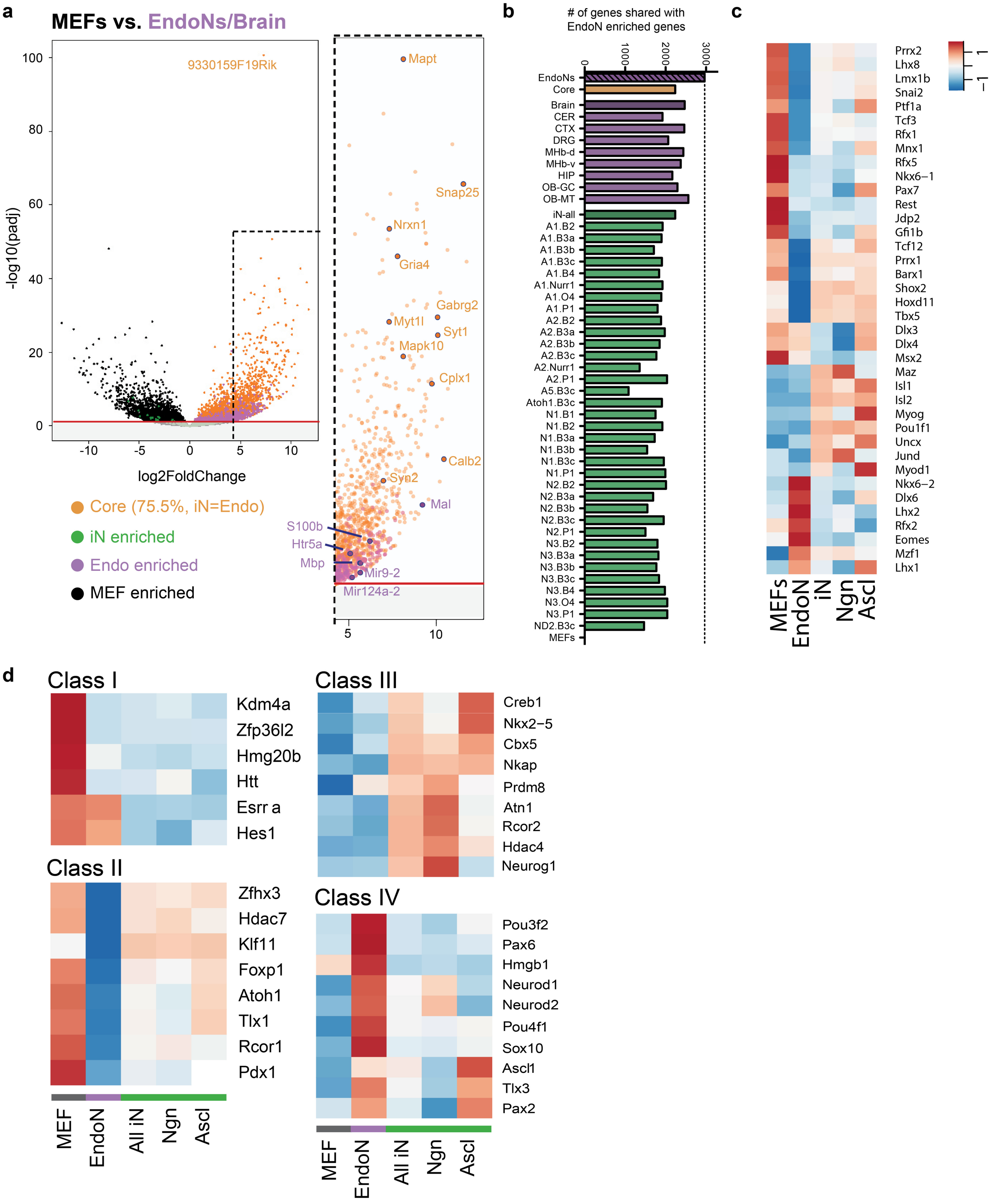


o

Population	Region	Strain	Age	Cell Type
Brain	whole brain	WT BALB/c	P66.5	Neural cells
CER	cerebellum	TauEGFP	P21	High intensity TauEGFP-positive cerebellar cells
CTX	cortex	TauEGFP	P21	High intensity TauEGFP-positive cortical cells
DRG	dorsal root ganglion	TauEGFP	P21	TauEGFP-positive dorsal root ganglion cells
MHb-d	habenula	Tac2-Cre/Ai9	P21	High intensity Tac2-Cre-positive habenula cells
MHb-v	habenula	Tac2-Cre/Ai9	P21	Low intensity Tac2-Cre-positive habenula cells
HIP	hippocampus	Tac2-Cre/Ai9	P21	Tac2-Cre-positive hippocampal cells
OB-GC	olfactory bulb	Pcdh21-Cre/Ai9	P21	Pcdh21-negative olfactory bulb cells
OB-MT	olfactory bulb	Pcdh21-Cre/Ai9	P21	Pcdh21-positive mitral and tufted cells

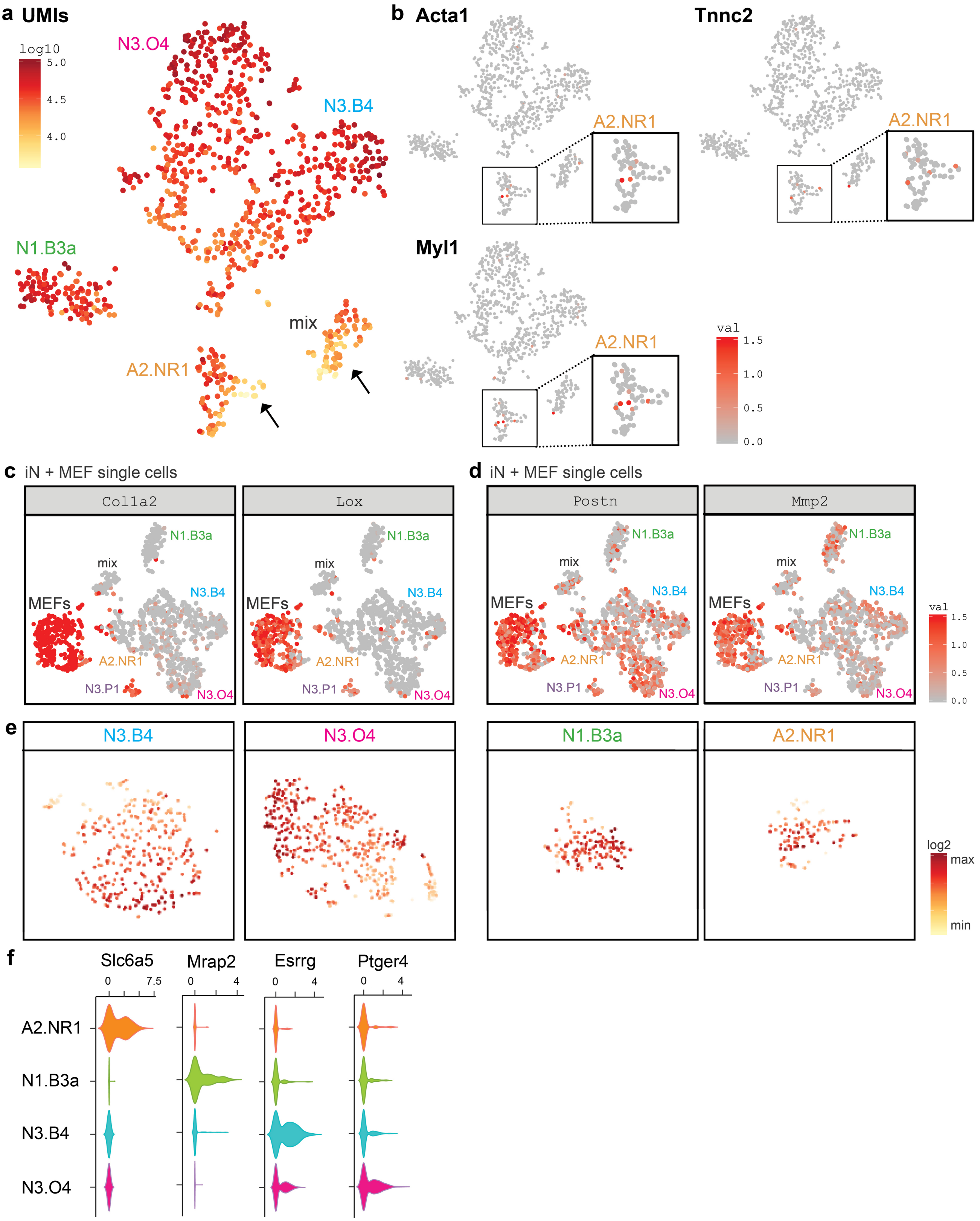
Extended Data Fig. 4 | See next page for caption.

Extended Data Fig. 4 | FACS, RNA-seq library preparation and characterization of iN cell and endogenous neuron populations. **a**, Representative immunofluorescence labelling of tau–EGFP$^+$ iN cell population (*Ascl2/Pou4f2*) on day 12 post-induction using neuronal antibodies TUJ1 and MAP2. Scale bars, 100 μm. *Pou4f2* is also known as *Brn3b*. **b**, Quantification of co-labelling of tau–eGFP and MAP2 in Tuj1$^+$ cells on day 12 post-induction calculated from various reprogramming transcription factor pairs. Data are presented as mean ± s.d. from $n = 4$ independent experiments and $n = 574$ cells. **c**, **d**, Representative FACS gates of an *Ascl2/Pou4f2* iN cell population (500,000 cells shown) (**c**) and a negative rtTA-only control (40,000 cells shown) (**d**) sorted on day 16 post-induction. Live tau–eGFP$^+$ cells were enriched by first gating DRAQ5$^+$ DAPI$^-$ cells, then collecting only those that were GFP$^+$. For *Ascl2/Pou4f2*, $n = 2$ independent experiments showed similar results, while for rtTA only, $n = 40$ independent experiments showed similar results. For all other iN cell populations, at least $n = 2$ independent experiments were performed to obtain biological replicates. **e**, Per cent of tau–eGFP$^+$ cells out of total number of cells collected post-FACS, presented as mean ± s.d. ($n = 4$ sorts, > 100 cells per sort). **f**, **g**, Correlation plots between aligned counts from single sequenced libraries of a *Neurog3/Pou3f2*-iN cell population generated from 10 ng versus 5 ng input RNA (**f**) and 10 ng versus 1 ng input RNA (**g**). *Pou3f2* is also known as *Brn2*. ***r***, Pearson correlation coefficient. **h**, Correlation plots between aligned counts from single sequenced libraries of a *Neurog3/Pou3f2* (10 ng input RNA) population and an *Ascl1/Pou3f2* (10 ng input RNA) population. **i**–**n**, Representative images taken while dissecting tissue from various brain regions of appropriate mouse reporter strains used to isolate specific endogenous cell-type populations used for RNA-seq: cerebellum (CER) (**i**), DRG (**j**), cortex (CTX) (**k**), olfactory bulb mitral and tufted cells (OB-MT) and olfactory bulb granule cells (OB-GC) (**l**), hippocampus (HIP) (**m**), and dorsal-medial habenula (MHb-d) and ventral-medial habenula (MHb-v) (**n**). $n = 2$ independent RNA-seq experiments. **o**, Characteristics of the endogenous neuron populations used for RNA-seq.



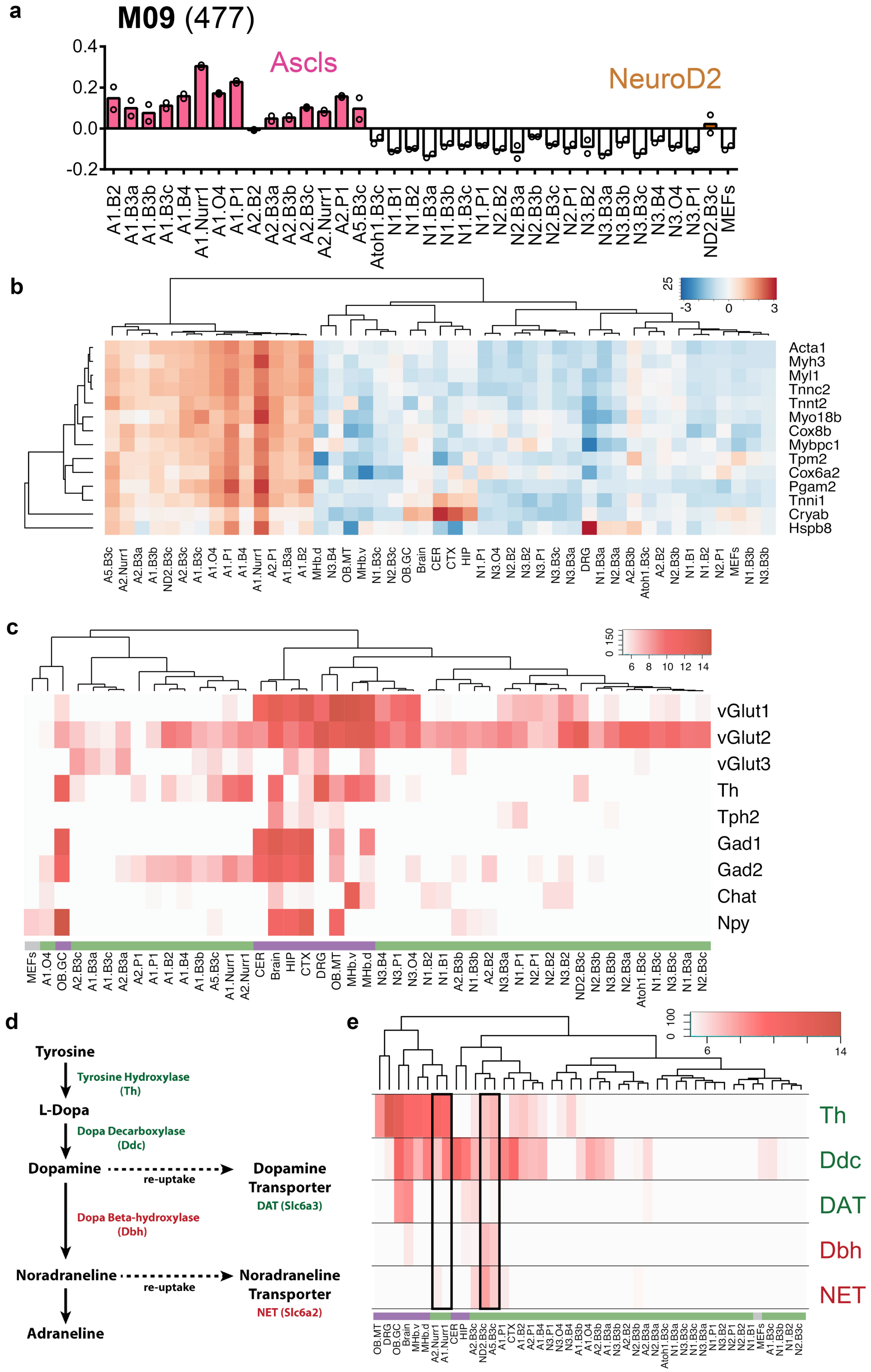
Extended Data Fig. 5 | See next page for caption.

Extended Data Fig. 5 | Upstream regulator analysis for core neuronal genes. a, Complete volcano plot of $\log_2$(fold change) versus –log(adjusted *P* value per gene) for MEFs (black) versus the pooled endogenous neuron and brain (endogenous neuron/brain) RNA-seq data. Genes enriched in MEFs and endogenous neuron/brain are plotted as negative and positive $\log_2$(fold change), respectively. Plotted are enriched core genes shared between iN cells and endogenous neuron/brain (orange, 75.5% of the significantly enriched endogenous neuron/brain genes), genes enriched in endogenous neuron/brain (purple, endo enriched), and genes enriched in iN cells (green, iN cell enriched). Red line, −log(0.05 *P*-adjusted value). Selected neural genes are labelled. **b**, Number of shared enriched genes between endogenous neurons and MEFs, individual endogenous neurons (purple) or iN cell (green) populations. Core genes (orange) are those shared collectively among iN cells and endogenous neurons. **c**, Heat map of expression of significant transcriptional regulators identified by HOMER only. Expression levels are defined as DESeq2 vsd-normalized RNA-seq counts with replicates averaged and scaled by row. **d**, Heat map of expression of significant class I–IV transcriptional regulators identified by IPA only. Class I, putative uniform neuronal repressor; Class II, putative non-uniform neuronal repressor; Class III, putative neuronal activator in iN cells; Class IV, putative neuronal activator in endogenous neurons. Expression levels are defined as DESeq2 vsd-normalized RNA-seq counts with groups averaged and scaled by row.



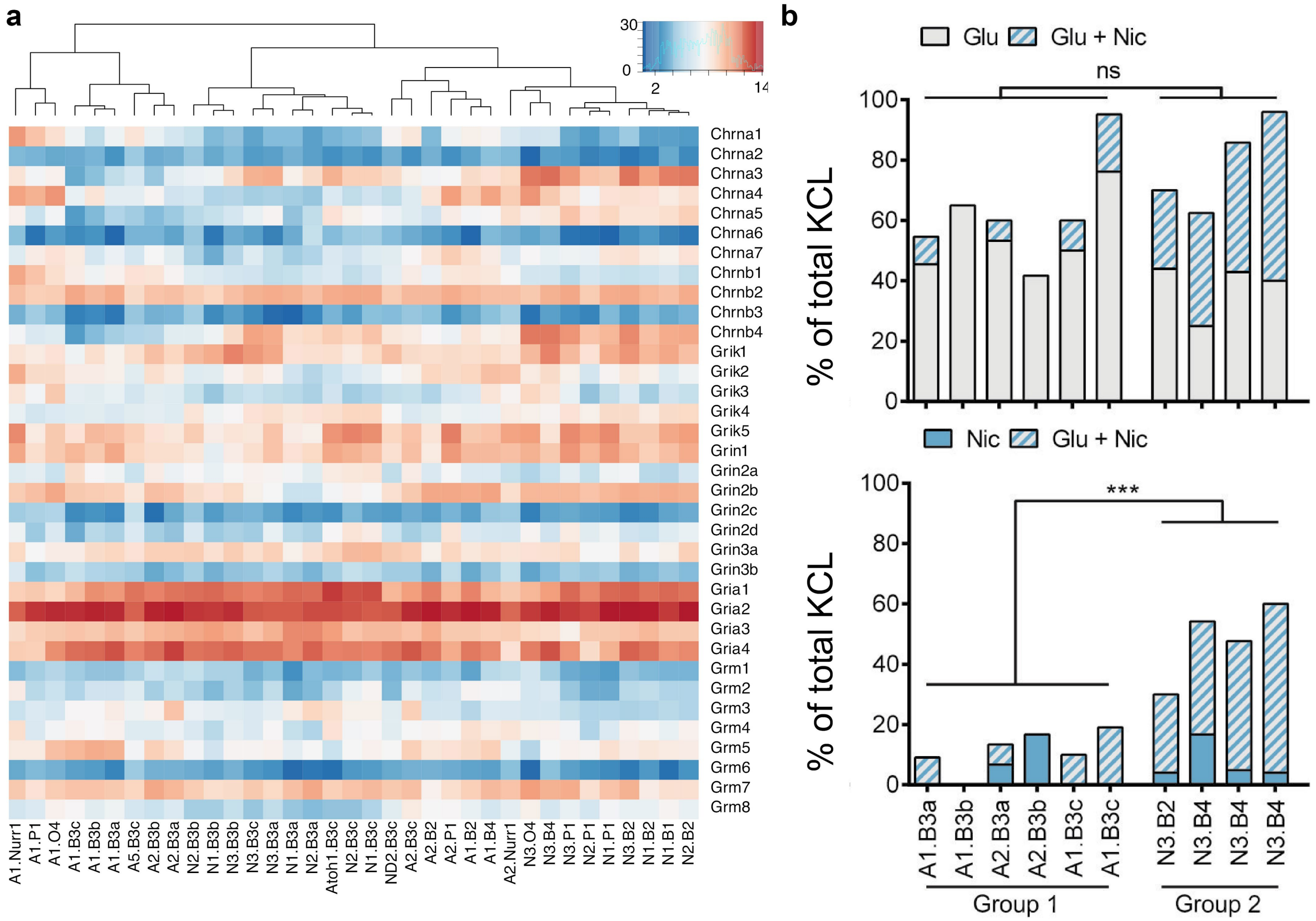
Extended Data Fig. 6 | See next page for caption.

Extended Data Fig. 6 | scRNA-seq analysis of iN cell populations. **a**, *t*-SNE projection of single cells collected from four iN cell populations, *Neurog3/Pou5f1* (N3.O4, $n = 415$ cells), *Neurog3/Pou3f4* (N3.B4, $n = 313$ cells), *Neurog1/Pou4f1* (N1.B3a, $n = 134$ cells) and *Ascl2/Nr4a2* (A2.NR1, $n = 90$ cells), coloured by log of UMI counts per cell. Arrows point to subpopulations of cells with low UMI, which includes a cluster composed of cells from each iN cell population. **b**, *t*-SNE projection of the same single cells shown in **a**, coloured by the log of UMI counts for the myogenic genes *Acta1*, *Tnnc2* and *Myl1*. Inset areas are magnified to highlight the small fraction of cells positive for the myogenic genes (3 out of 90, threshold set at log(UMI counts) > 1) in the *Ascl2/Nr4a2* (A2.NR1) population. The three myogenic genes plotted were those identified previously[19] that were not highly expressed in any of our endogenous neuron populations. **c**, *t*-SNE projection of single cells collected from MEFs and five iN cell populations: *Neurog3/Pou5f1* (N3.O4), *Neurog3/Pou3f4* (N3.B4), *Neurog1/Pou4f1* (N1.B3a), *Ascl2/Nr4a2* (A1.NR1) and *Neurog3/Pou1f1* (N3.P1). Cells are coloured by the log of UMI counts for genes *Col1a2* and *Lox*, which represent MEF genes (10 out of 15 genes) that are highly expressed in the majority of the MEF population and in a small fraction of cells in the iN cell populations. Fifteen MEF genes were selected, based on the top genes enriched in MEFs compared to endogenous neuron/brain according to population RNA-seq that were not also expressed in endogenous neuronal single cells (data not shown). **d**, *t*-SNE projection of the same single cells as shown in **c**, coloured by the log of UMI counts for the genes *Postn* and *Mmp2*, which represent MEF genes (5 out of 15 genes) that are highly expressed in the majority of the MEF population and in a large fraction of cells in the iN cell populations. **e**, *t*-SNE projections of single cells coloured by log of UMI counts per cell for each of the individual iN cell populations sequenced: *Neurog3/Pou3f4* (N3.B4), *Neurog3/Pou5f1* (N3.O4), *Neurog1/Pou4f1* (N1.B3a) and *Ascl2/Nr4a2* (A1.NR1). The number of cells for each transcription factor combination is the same as in **a**. **f**, Expression of receptors and transmembrane proteins among the top 20 differentially expressed genes in each transcription factor pair relative to all other combinations, plotted as a simplified violin plot. One representative gene shown for each transcription factor pair.

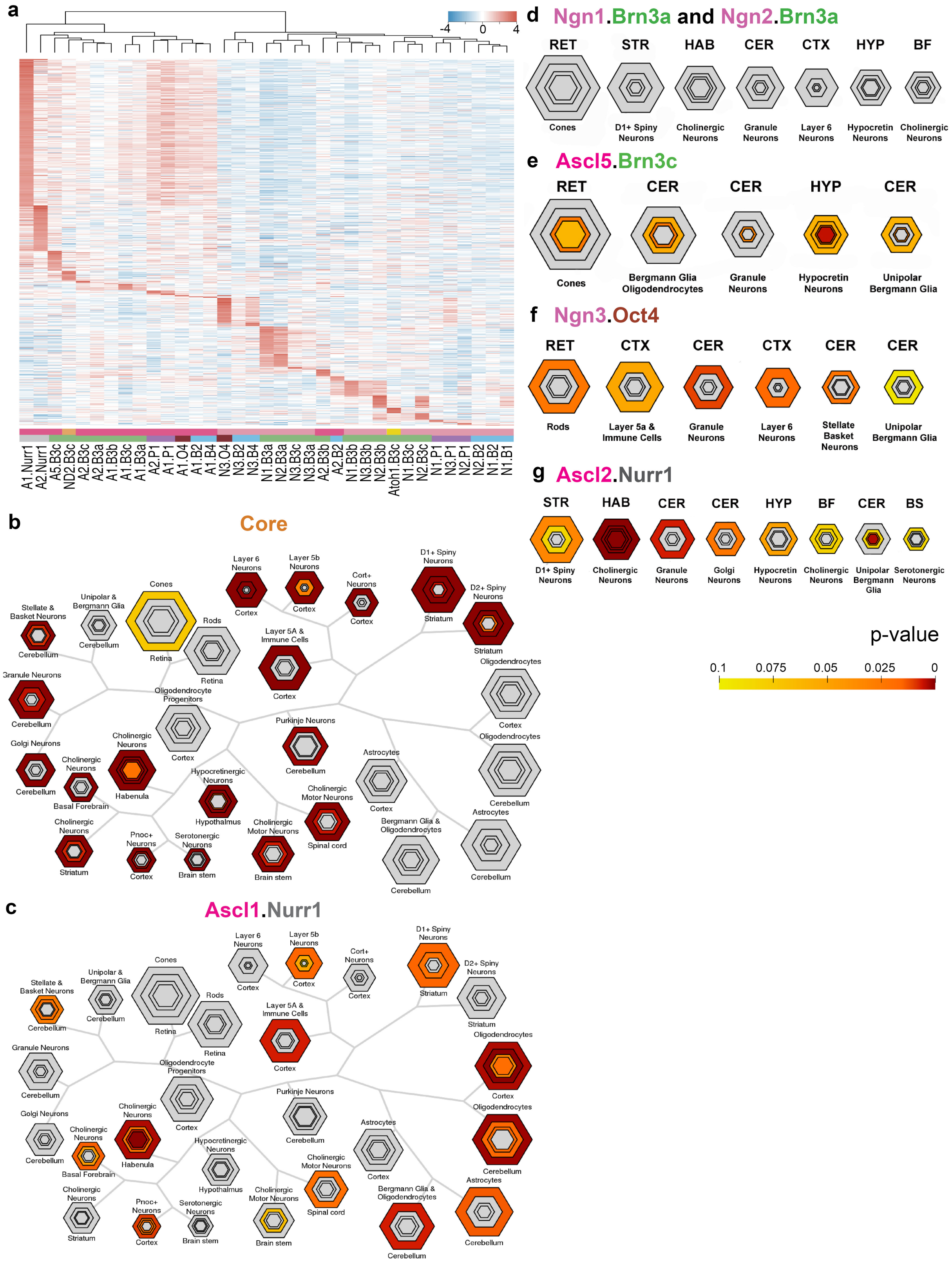


Extended Data Fig. 7 | See next page for caption.

Extended Data Fig. 7 | Diversity across iN cell populations. a, WGCNA module eigengene expression of the 35 iN cell populations (in duplicate) shown as bar plots of average module eigengene expression for module 09 (M09, $n = 477$ genes) correlated with bHLH subclasses. Colours highlight iN cells populations generated with the Ascl family of bHLH factors or an iN cell combination generated with the bHLH factor, *Neurod2*. **b**, Heat map of expression of myogenic genes reflects higher levels of expression in iN cell populations derived with the Ascl family of reprogramming factors compared to the Neurog family. The myogenic gene list is as described[19]. Expression levels are defined as DESeq2 vsd-normalized RNA-seq counts with replicates averaged and scaled by row. The dendrogram represents hierarchical clustering based on correlation distance. **c**, Heat map of expression of select neurotransmitter-associated genes. Expression levels in iN cell (green), endogenous neuron/brain (purple) and MEF populations (grey) are defined as DESeq2 vsd-normalized RNA-seq counts with replicates averaged. Dendrogram represents hierarchical clustering based on correlation distance. **d**, Schematic of dopamine and noradrenaline biosynthesis pathway. **e**, Heat map of expression of genes involved in dopamine and noradrenaline biosynthesis and re-uptake across all iN cell (green), endogenous neurons (purple) and MEF (grey) populations. Expression patterns for populations generated with *Ascl1/Nr4a2*, *Ascl2/Nr4a2*, *Ascl5/Pou4f3* and *Neurod2/Pou4f3* are outlined with a black frame. Expression levels are defined as DESeq2 vsd-normalized RNA-seq counts with replicates averaged. Dendrogram represents hierarchical clustering based on correlation distance.

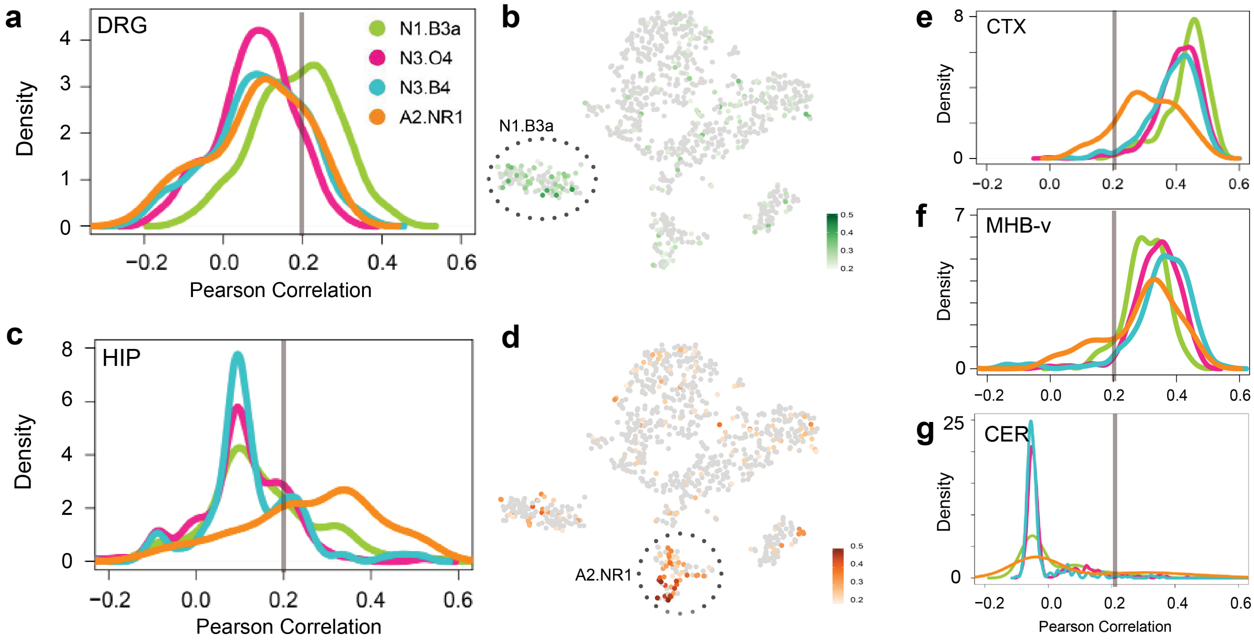


Extended Data Fig. 8 | Diverse glutamate and nicotine responses of iN cell populations. a, Heat map of expression of glutamate and nicotinic acetylcholine receptor subunit genes across all iN cell populations. Expression levels are defined as DESeq2 vsd-normalized RNA-seq counts with replicates averaged. Dendrogram represents hierarchical clustering based on correlation distance. **b**, Percentages of glutamate- and nicotine-responsive cells out of total KCl-responsive cells in each individual iN cell population ($n = 218$ total cells). Group 1 ($n = 6$ independent experiments) and group 2 ($n = 4$) comprise iN cell populations with the lowest and highest overall expression of nicotinic acetylcholine receptors, respectively. ***$P = 0.0004$; ns, not significant (unpaired Student's t-test). Data are mean ± s.d.



Extended Data Fig. 9 | See next page for caption.

Extended Data Fig. 9 | Comparison of iN cell and endogenous neuron populations. **a**, Heat map of expression of uniquely enriched genes in individual iN cell populations as defined by genes significantly enriched (*P*-adjusted value < 0.05) in each iN cell population versus all other iN cell populations and MEFs determined by DESeq2. Expression levels are defined as DESeq2 vsd-normalized RNA-seq counts with replicates averaged and scaled by row. Dendrogram represents hierarchical clustering based on Euclidean distance. **b**–**g**, Overlap of gene lists with a particular cell type or region for which data are currently available were identified by Fisher's exact test (two-sided) with Benjamini–Hochberg correction using CSEA. Concentric, hexagonal plots represent each cell type or region. The sizes of the hexagons are scaled to the number of specifically enriched transcripts at set stringency thresholds with the innermost hexagon representing the most unique genes. Hexagons are colour coded by the *P* values of the Fisher's exact test. RET, retina; HYP, hypothalamus; STR, striatum; HAB, habenula; BF, basal forebrain; BS, brainstem. **b**–**d**, CSEA of the core genes (enriched genes shared between iN cell and endogenous neuron/brain populations, $n = 2{,}239$ genes) (**b**) and uniquely enriched genes of iN cell populations *Ascl1/Nr4a2* (A1.Nurr1, $n = 282$ genes) (**c**) and *Neurog1/Pou4f1* and *Neurog2/Pou4f1* (N1/N2.B3a, combined $n = 93$ genes total) (**d**). Uniquely enriched genes were defined in the same manner as in **a**. **e**–**g**, Modified CSEA visualization of uniquely enriched genes of individual iN cell populations: *Ascl5/Pou4f3* (A5.B3c, $n = 46$ genes) (**e**), *Neurog3/Pou5f1* (N3.O4, $n = 51$ genes) (**f**) and *Ascl2/Nr4a2* (A2.NR1, $n = 101$ genes) (**g**). Uniquely enriched genes were defined in the same manner as in **a**.



Extended Data Fig. 10 | Comparison of single-cell iN cell and endogenous neuron populations. a, Pearson correlation values between individual single cells and bulk DRG plotted as kernel density distributions for each transcription factor pair, and colour-coded accordingly. To generate Pearson correlation values between endogenous populations and single cells, unique genes for each endogenous population ($n = 1$ population in duplicate biological samples) were defined using DESeq2 as the top 100 significant genes that were ranked by highest fold change when compared to all other endogenous populations ($n = 5$ in duplicate, $n = 2$ in triplicate biological samples). The expression level of these unique genes in their respective endogenous population was correlated with each single-cell for genes that were found in filtered gene-barcode matrices. Pearson correlation values were plotted as kernel density estimations to represent the distribution of single cells for each iN cell population: *Neurog1/Pou4f1* (N1.B3a, $n = 134$ cells, green), *Neurog3/Pou5f1* (N3.O4, $n = 415$ cells, pink), *Neurog3/Pou3f4* (N3.B4, $n = 313$ cells, blue) and *Ascl2/Nr4a2* (A2.NR1, 90 cells, orange). **b**, *t*-SNE projections of 952 single cells coloured by their correlation with bulk DRG. The *Neurog1/Pou4f1* pair exhibits enrichment of highly correlated cells. **c**, Pearson correlation values between individual single cells ($n = 952$ cells) and bulk HIP plotted as kernel density distributions for each combination and colour coded accordingly. **d**, *t*-SNE projections of 952 single cells coloured by their correlation with bulk HIP. The *Ascl2/Nr4a2* pair exhibits enrichment of highly correlated cells. **e**–**g**, Pearson correlation values between individual single cells and bulk CTX (**e**), MHb-v (**f**) and CER (**g**) plotted as kernel density distributions for each combination, and colour-coded accordingly.

https://doi.org/10.1038/s41586-018-0079-1

Catalytic activation of β-arrestin by GPCRs

Kelsie Eichel[1,2], Damien Jullié[1,2], Benjamin Barsi-Rhyne[1,2], Naomi R. Latorraca[3,4,5,6], Matthieu Masureel[5],
Jean-Baptiste Sibarita[7,8], Ron O. Dror[3,4,5,6] & Mark von Zastrow[1,2]*

β-arrestins are critical regulator and transducer proteins for G-protein-coupled receptors (GPCRs). β-arrestin is widely believed to be activated by forming a stable and stoichiometric GPCR–β-arrestin scaffold complex, which requires and is driven by the phosphorylated tail of the GPCR. Here we demonstrate a distinct and additional mechanism of β-arrestin activation that does not require stable GPCR–β-arrestin scaffolding or the GPCR tail. Instead, it occurs through transient engagement of the GPCR core, which destabilizes a conserved inter-domain charge network in β-arrestin. This promotes capture of β-arrestin at the plasma membrane and its accumulation in clathrin-coated endocytic structures (CCSs) after dissociation from the GPCR, requiring a series of interactions with membrane phosphoinositides and CCS-lattice proteins. β-arrestin clustering in CCSs in the absence of the upstream activating GPCR is associated with a β-arrestin-dependent component of the cellular ERK (extracellular signal-regulated kinase) response. These results delineate a discrete mechanism of cellular β-arrestin function that is activated catalytically by GPCRs.

GPCRs, the largest family of signalling receptors, regulate essentially every physiological process and comprise an important class of drug targets[1–4]. GPCR-mediated signalling and regulatory events occur primarily through interactions of the receptor with two classes of transducer protein, heterotrimeric G proteins and β-arrestins. β-arrestins were discovered through their ability to prevent coupling of G proteins to GPCRs and are now known to support additional functions, including endocytosis of GPCRs, mediated by clathrin-coated structures, and downstream signalling, mediated by MAP (mitogen-activated protein) kinase cascades[5,6]. A long-standing view is that all of these functions occur from a stable and stoichiometric GPCR–β-arrestin complex, the formation of which requires β-arrestin binding to the phosphorylated GPCR tail[7,8]. There is emerging evidence that GPCR–β-arrestin complexes can vary in structure but, nevertheless, present concepts of cellular β-arrestin function require the formation of a GPCR–β-arrestin complex driven largely by the phosphorylated GPCR tail[9–13].

Recently, β-arrestin-2 was found to mediate MAP kinase signalling by accumulating in CCSs in response to ligand-dependent activation of the β1-adrenergic GPCR (β1AR) but without co-accumulation of the β1AR. This ability of β-arrestin-2 to operate separately from its activating GPCR is not consistent with the present mechanistic understanding and remains unexplained. Here we show that such 'action at a distance' behaviour is widespread. We delineate a distinct, GPCR tail-independent mechanism of cellular β-arrestin activation in which transient engagement of the GPCR acts catalytically.

Separate trafficking of β-arrestin

We verified separate trafficking of β-arrestin-2 in HEK 293 cells co-expressing recombinant β1ARs using total internal reflection fluorescence (TIRF) microscopy. The β-adrenergic agonist isoproterenol produced rapid and robust accumulation of β-arrestin-2 in CCSs without detectable co-accumulation of β1AR (Fig. 1a, b). Nevertheless, trafficking of β-arrestin-2 to CCSs was dependent on ligand-induced

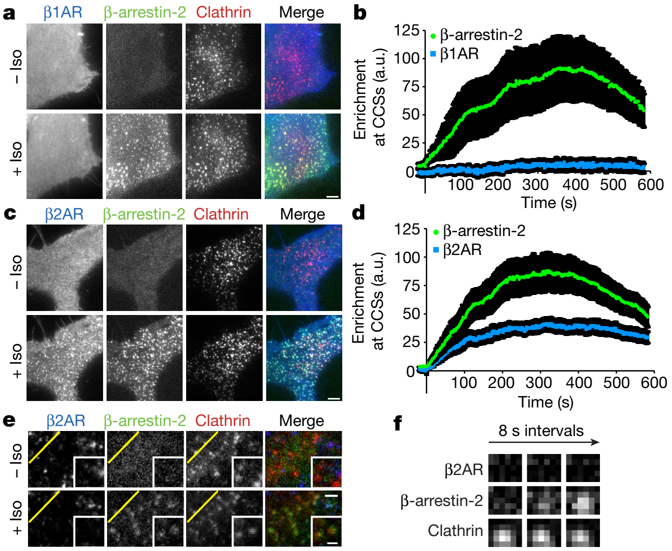


Fig. 1 | Discrete mode of GPCR-activated cellular β-arrestin trafficking is broadly conserved. a–d, Live-cell TIRF microscopy images showing Flag–β1AR (blue) (**a**) or Flag–β2AR (blue) (**c**), β-arrestin-2–GFP (green) and clathrin-light-chain–DsRed (red) before and after treatment with 10 μM isoproterenol (Iso). **b**, **d**, Mean enrichment of Flag–β1AR (**b**) and Flag–β2AR (**d**) at CCSs after treatment with 10 μM isoproterenol. $n = 14$ and 15 cells, respectively, from three independent experiments; data shown as mean ± s.e.m. **e**, Live-cell TIRF microscopy images of HEK 293 cells co-expressing immobilized super-ecliptic pHluorin–β2AR (blue), β-arrestin-2–mApple (green) and clathrin light chain–TagBFP (red) before and after treatment with 10 μM isoproterenol. **f**, Time-lapse images of individual pre-existing CCSs from **e**. Scale bars, 5 μm. **a**, **c**, **e** and **f** show representative images from three independent experiments.

---

[1]Department of Cellular and Molecular Pharmacology, University of California, San Francisco School of Medicine, San Francisco, CA, USA. [2]Department of Psychiatry, University of California, San Francisco School of Medicine, San Francisco, CA, USA. [3]Biophysics Program, Stanford University, Stanford, CA, USA. [4]Department of Computer Science, Stanford University, Stanford, CA, USA. [5]Department of Molecular and Cellular Physiology, Stanford University School of Medicine, Stanford, CA, USA. [6]Institute for Computational and Mathematical Engineering, Stanford University, Stanford, CA, USA. [7]Interdisciplinary Institute for Neuroscience, UMR 5297, Centre National de la Recherche Scientifique, Bordeaux, France. [8]Interdisciplinary Institute for Neuroscience, University of Bordeaux, Bordeaux, France. *e-mail: mark.vonzastrow@ucsf.edu

β1AR activation, as it was greatly reduced by the β1-selective antagonist CGP 20712A (Extended Data Fig. 1a) or in HEK 293 cells that did not express recombinant β1ARs (Extended Data Fig. 1b). In H9c2 cells, which natively express β1ARs at higher levels[14], either isoproterenol or the β1-selective agonist dobutamine activated β-arrestin trafficking through endogenous receptors (Extended Data Fig. 1c) and the β1-selective antagonist CGP 20712A inhibited this (Extended Data Fig. 1d). We also found that β-arrestin-1 (also known as arrestin-2) is capable of separate accumulation in CCSs (Extended Data Fig. 1e–g), establishing generality across β-arrestin isoforms.

The β2-adrenergic receptor (β2AR) is a homologous GPCR that co-accumulates with β-arrestin in CCSs[15–17]. We verified this by TIRF microscopy in HEK 293 cells expressing Flag-tagged β2ARs at levels around tenfold higher than endogenous levels (Fig. 1c, d, Extended Data Fig. 1h). We considered the possibility that β2ARs can also activate β-arrestin trafficking separately from the receptor, but that this capacity is obscured by the natural tendency of β2ARs to co-traffic. Indeed, when laterally immobilized to prevent receptor accumulation in CCSs[18,19], Flag–β2AR still promoted rapid accumulation of β-arrestin-2 in CCSs (Fig. 1e, f, Extended Data Fig. 1i, j). Furthermore, when not immobilized, both β1AR and β2AR produced super-stoichiometric accumulation of β-arrestin-2 in CCSs relative to receptor (around 28- and 4-fold, respectively; Extended Data Fig. 1k–m). Moreover, a limited survey of family A GPCRs suggested that such differential activation of β-arrestin trafficking is broadly conserved across GPCRs as well as across β-arrestin isoforms (Extended Data Fig. 2).

Activation by GPCR core interaction

Canonical GPCR–β-arrestin scaffolding requires the receptor cytoplasmic tail and is driven by phosphorylation of the tail[20–23]. By contrast, the β1AR cytoplasmic tail was not required to activate β-arrestin trafficking. β-arrestin accumulation in CCSs was unaffected by nearly complete removal of the GPCR cytoplasmic tail (415T, Extended Data Fig. 3a, b). This was also true for β2ARs after a similarly extensive tail truncation (341T, Fig. 2a, b), as well as after a less extensive truncation (365T) that was previously shown to abrogate β2AR–β-arrestin scaffold complex formation[9] (Extended Data Fig. 3c–e).

In principle, GPCRs could promote tail-independent trafficking of β-arrestin through activation of a downstream G-protein-linked signalling pathway. We found that this was not the case, as GPCR-activated β-arrestin-2 accumulation in CCSs did not depend on selectivity for G protein coupling, and trafficking activated by the G_i-coupled D2 dopamine receptor (DRD2) was unaffected by pertussis toxin (Extended Data Fig. 3f–h), consistent with a recent report[24]. Further, receptor-independent activation of adenylyl cyclase using forskolin did not promote (or block) β-arrestin-2 trafficking (Extended Data Fig. 3i–k). Rather, the discrete β-arrestin trafficking behaviour required direct binding of the ligand-activated GPCR, because a mutant β-arrestin-2 that is unable to bind GPCRs[25] did not accumulate in CCSs (Extended Data Table 1; Extended Data Fig. 3l–n). β-arrestin has been shown to engage GPCRs weakly through a ligand-dependent interaction that does not require the phosphorylated receptor tail[26]. We hypothesized that transient binding is sufficient to activate β-arrestin trafficking and is mediated by the GPCR core. To test this, we focused on DRD2 because it has a short cytoplasmic tail and yet robustly activates β-arrestin trafficking to CCSs (Extended Data Fig. 2g–i). A mutation in the core of DRD2 that specifically disrupts coupling of the receptor to β-arrestin (DRD2(G prot)[27]) prevented ligand-dependent stimulation of β-arrestin trafficking to CCSs (Fig. 2c–e), establishing a critical role for the interaction with the receptor core.

We next investigated how engagement with the receptor core affects the activation of β-arrestin trafficking by focusing on a polar region in β-arrestin that is located proximal to the conserved finger loop and is thought to directly contact the ligand-activated GPCR core[10,21]. We identified three charged residues in this ‘finger-loop-proximal’ region that produce constitutive β-arrestin-2 accumulation in CCSs when mutated to alanine (R77A, K78A, D79A; Fig. 3a, b). When analysed using a statistical metric of β-arrestin clustering validated against a previously described constitutively active β-arrestin construct (polar core mutant[28]; Extended Data Table 1, Extended Data Fig. 4a, b), these finger-loop-proximal alanine substitutions produced a comparably strong constitutive activation phenotype (Extended Data Fig. 4b). The finger-loop-proximal charge mutations also increased ligand-independent interaction of β-arrestin with the clathrin-associated adaptor protein-2 (AP-2; Extended Data Fig. 4c–e), providing biochemical evidence for constitutive activation of β-arrestin. Individual substitution of finger-loop-proximal charged residues was sufficient to produce ligand-independent accumulation of β-arrestin-2 in CCSs, with the K78A mutation having the largest effect (Extended Data Fig. 4b, f–h). Further, mutating K78 to arginine (K78R) rather than alanine did not produce constitutive activation (Extended Data Fig. 4b, i). Together, these results suggest that K78 stabilizes β-arrestin in its inactive (cytoplasmic) state through a charge interaction positioned at the GPCR core-binding interface.

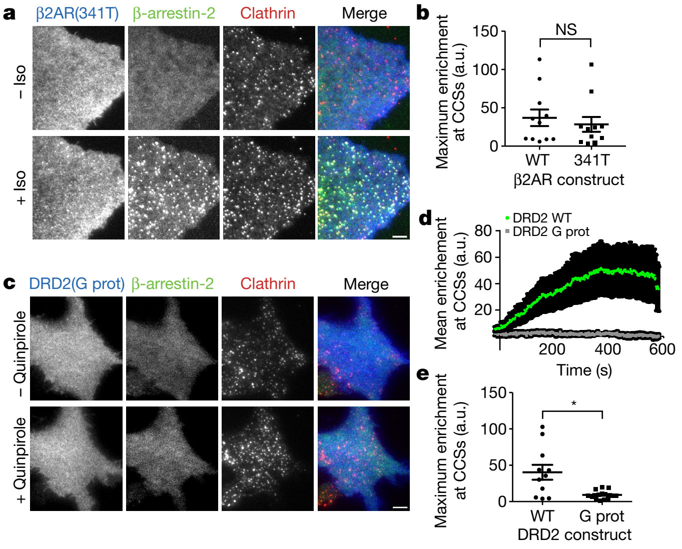


Fig. 2 | Activation of β-arrestin trafficking requires the GPCR core but not the GPCR cytoplasmic tail. a, Live-cell TIRF microscopy images of COS-1 cells co-expressing Flag–β2AR truncated at residue 341 (β2AR(341T), blue), β-arrestin-2–GFP (green) and clathrin light chain–DsRed (red) before and after treatment with 10 μM isoproterenol. **b**, Maximum β-arrestin-2–GFP enrichment at CCSs in cells treated with 10 μM isoproterenol and co-expressing the indicated Flag–β2AR construct. $n = 11$ cells from three independent experiments; $P = 0.5634$, two-tailed unpaired t-test. NS, not significant. **c**, Live-cell TIRF microscopy images showing Flag–DRD2 G protein-biased mutant (DRD2(G prot), blue), β-arrestin-2–GFP (green) and clathrin light chain–DsRed (red) before and after treatment with 10 μM quinpirole. **d**, Mean (±s.e.m.) enrichment of β-arrestin-2–GFP in CCSs as shown in **c**. Green, wild-type (WT) Flag–DRD2; grey, Flag–DRD2(G prot). **e**, Maximum enrichment of β-arrestin-2–GFP in CCSs of cells as shown in **c** treated with 10 μM quinpirole. In **d** and **e**, $n = 11$ (wild-type) and 14 (G protein-biased) cells from three independent experiments; $P = 0.013$, two-tailed unpaired t-test using Welch's correction. **a** and **c** show representative images from three independent experiments. Scatter plots show overlay of mean and s.e.m. Scale bars, 5 μm. *$P < 0.05$.

To investigate how these residues may affect β-arrestin function, we focused on β-arrestin-1, because more structural data are available for this isoform than for β-arrestin-2 and mutating the corresponding residues in β-arrestin-1 produced constitutive accumulation in CCSs (Fig. 3c). Crystal structures of putative inactive (Extended Data Fig. 5a left, middle) and active (Extended Data Fig. 5a right) forms of β-arrestin-1 reveal that the finger-loop-proximal charged residues are located within an extensive network of polar residues spanning the N and C domains of β-arrestin. In both inactive and active structures, R76 may interact with D78 (corresponding to R77 and D79 in

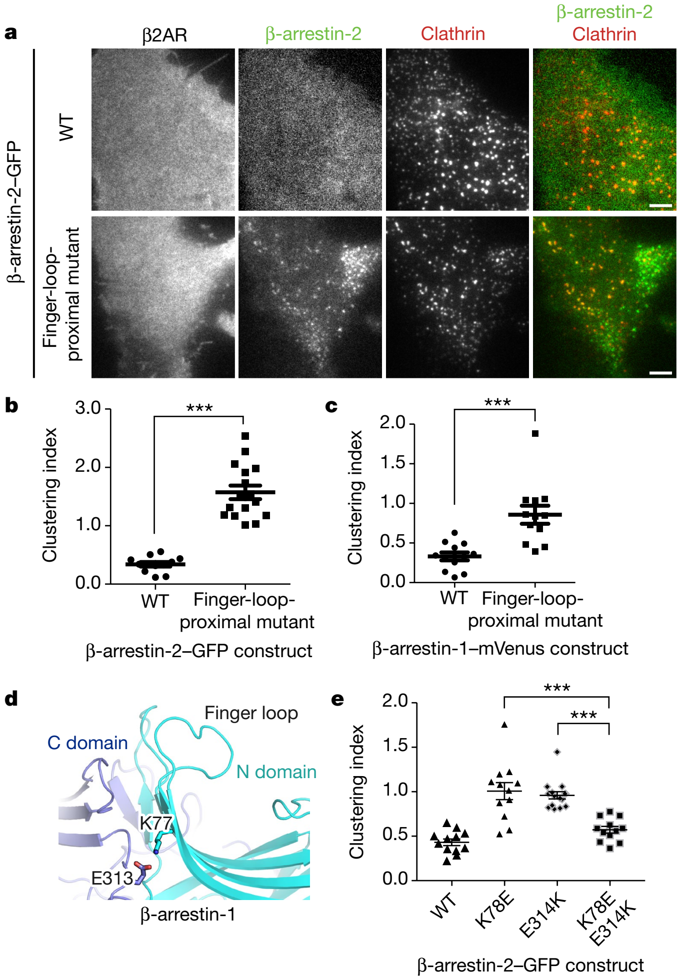


Fig. 3 | β-arrestin activation is inhibited by a polar network in a region proximal to the β-arrestin finger loop. a, Representative live-cell TIRF microscopy images (from three independent experiments) showing Flag–β2AR (blue), clathrin light chain–DsRed (red) and wild-type (top, green) or finger-loop-proximal-mutant (bottom, green) β-arrestin-2–GFP without agonist treatment. **b**, **c**, Clustering index measuring constitutive activation of the indicated β-arrestin-2–GFP (**b**) or β-arrestin-1–mVenus (**c**) constructs without agonist treatment. $n = 12$ cells from three independent experiments; $P < 0.0001$ and 0.0008, respectively, two-tailed unpaired *t*-test. **d**, Snapshot from molecular dynamics simulations of inactive-state β-arrestin-1 in which K77 and E313 occasionally form a stable salt bridge. This salt bridge formed 6% of the time in inactive-state simulations (six simulations, total duration = 26.7 μs); it may form more frequently on longer timescales. It formed in only a few frames of active-state simulations (0.2% of the time across six simulations, total duration = 29.3 μs). **e**, Clustering index of the indicated β-arrestin-2–GFP construct without agonist treatment. Statistics were calculated using a two-tailed unpaired *t*-test. For K78E, $n = 12$ cells from three independent experiments, $P = 0.0003$; for E314K, $n = 12$ cells from three independent experiments, $P < 0.0001$. Scatter plots show overlay of mean and s.e.m. Scale bars, 5 μm. ***$P < 0.001$.

β-arrestin-2) but there is no obvious binding partner for K77 (K78 in β-arrestin-2), even though point mutation of this residue produced the strongest constitutive phenotype. We used molecular dynamics simulations to identify acidic residues that might interact with K77. In inactive-state simulations, K77 sometimes formed an intramolecular salt bridge with E313 in the C domain, but this interaction rarely formed in active-state simulations (Fig. 3d, Extended Data Fig. 5b). This interaction may have been overlooked previously because E313 interacts with R188 on another β-arrestin molecule in the crystal lattice, probably owing to lattice packing (Extended Data Fig. 5c). Both K77 and E313 are conserved in β-arrestin-2 (K78 and E314, Extended Data Fig. 5d) and, in both β-arrestin isoforms, separation of N and C domains is thought to accompany β-arrestin activation[20,29,30]. Therefore we hypothesized that K77/78 and E313/314 occasionally form a salt bridge that stabilizes the inactive state of β-arrestin by favouring tighter interactions between the N and C domains. Charge mutation of E314 (E314K) also produced ligand-independent accumulation of β-arrestin-2 in CCSs (Fig. 3e). Further, mutating both residues (K78E/E314K) to restore the putative ionic interaction by charge swap reversed this constitutive phenotype (Fig. 3e). Together, these findings suggest that the finger-loop-proximal charged residues function as part of an inter-domain interaction network that maintains β-arrestin in its inactive cytoplasmic form and is destabilized by interaction with the GPCR core.

Capture after GPCR dissociation

For β-arrestin to traffic to CCSs after dissociating from its upstream activating GPCR, additional partner(s) must engage and stabilize β-arrestin at the plasma membrane. We focused on several known candidates (Extended Data Table 1, Extended Data Fig. 6a, b). Mutating a conserved phosphoinositide binding determinant in the β-arrestin C domain that was previously mapped and implicated in β-arrestin trafficking[31] prevented ligand-induced accumulation of β-arrestin-2 at the plasma membrane and in CCSs (lipid mutant, Fig. 4a–i, Extended Data Fig. 6c, d). We demonstrated biochemically that binding of phosphatidylinositol 4,5-bisphosphate (PtdIns(4,5)P_2) to this determinant is sufficient to partition β-arrestin out of the solution phase (Extended Data Fig. 6e, f). This phosphoinositide-binding determinant is specific to β-arrestins, but a lipid-anchoring region in the C domain was recently identified in visual arrestin (also known as arrestin-1)[32]. Mutating the homologous residues in β-arrestin-2 did not prevent accumulation at the plasma membrane or in CCSs (Extended Data Table 1, Extended Data Fig. 6g, h). Mutating previously identified clathrin[33] and AP-2[34] binding determinants (CCS mutant) in the C terminus of β-arrestin prevented β-arrestin-2 from accumulating in CCSs without blocking accumulation at the plasma membrane (Extended Data Table 1, Fig. 4j–l, Extended Data Fig. 6i). Mutating the phosphoinositide binding determinant together with clathrin and AP-2 binding determinants blocked β-arrestin-2 accumulation at the plasma membrane altogether, the same phenotype that resulted from mutating only the phosphoinositide binding determinant (lipid and CCS mutant, Fig. 4m–o, Extended Data Fig. 6j). Similar results were obtained using the β1AR rather than β2AR as the activating GPCR (Extended Data Fig. 6k–ac). Further, mutating the phosphoinositide binding determinant prevented the constitutive trafficking phenotype produced by finger-loop-proximal charge mutations in β-arrestin-2 (Extended Data Fig. 6ad, ae). These results indicate that β-arrestin is stabilized at the plasma membrane after dissociating from its activating GPCR through a series of non-GPCR interactions.

Whereas the phosphoinositide binding determinant was essential for β1AR- or β2AR-dependent trafficking of β-arrestin-2 to CCSs, it was not required for accumulation of β-arrestin-2 in CCSs produced by a chimaeric construct of β2AR containing the tail region of the V2 vasopressin receptor (β2AR–V2R chimaera), a GPCR that is known to form a highly stable tail-dependent GPCR–β-arrestin scaffold complex[35] (Extended Data Fig. 7a–c). Independently verifying this difference, depleting PtdIns(4,5)P_2 from the plasma membrane using phenylarsine oxide (PAO)[36] blocked the ability of β2AR but not that of the β2AR–V2R chimaera to activate trafficking of wild-type β-arrestin-2 to CCSs (Extended Data Fig. 7d–g). Thus, phosphoinositide binding to β-arrestin appears to be required specifically for trafficking of β-arrestin to CCSs after it dissociates from its upstream activating GPCR, but not for trafficking to CCSs while bound in a sufficiently stable GPCR–β-arrestin scaffold complex.

Dynamics in the plasma membrane

We next investigated the discrete trafficking mechanism using single-particle tracking–photoactivated localization microscopy

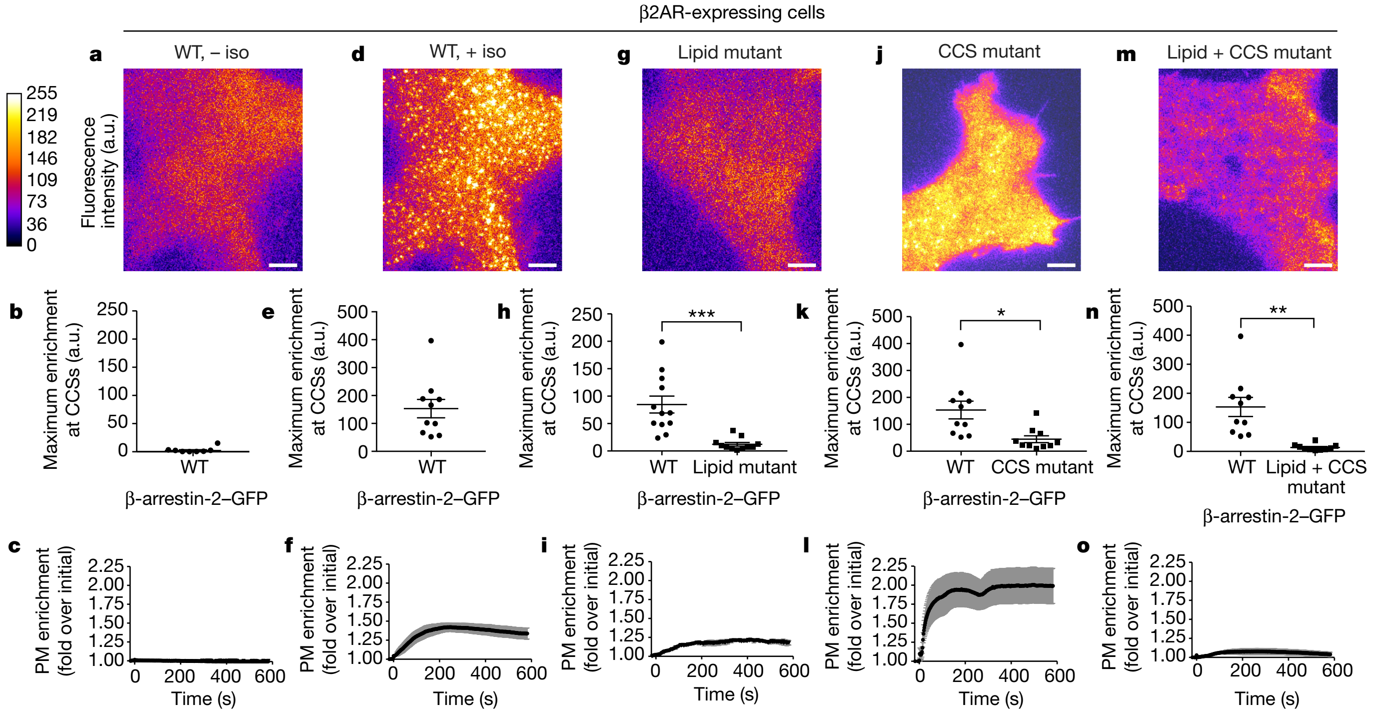


Fig. 4 | Phosphoinositide binding is required to capture β-arrestin at the plasma membrane after GPCR dissociation. **a**–**o**, Representative live-cell TIRF microscopy images (**a**, **d**, **g**, **j**, **m**; three independent experiments) showing Flag–β2AR, the indicated β-arrestin-2–GFP construct and clathrin light chain–DsRed. Shown are β-arrestin images false-coloured to indicate fluorescence intensity (**a**, **d**, **g**, **j**, **m**), maximum fluorescence enrichment at CCSs (**b**, **e**, **h**, **k**, **n**) and normalized mean plasma membrane β-arrestin-2–GFP fluorescence (**c**, **f**, **i**, **l**, **o**), respectively, from cells co-expressing Flag–β2ARs without isoproterenol treatment (**a**–**c**), and the following β-arrestin-2–GFP constructs with 10 μM isoproterenol treatment: wild-type (**d**–**f**), lipid mutant (**g**–**i**), CCS mutant (**j**–**l**), and CCS and lipid double mutant (**m**–**o**). Wild-type β-arrestin-2–GFP maximum enrichment in **k** and **n** is replotted from data shown in **e**. PM, plasma membrane. Statistics were calculated using a two-tailed unpaired *t*-test with Welch's correction. In **h**, $n = 12$ and 11 cells, respectively, from three independent experiments, $P = 0.0006$. In **k**, $n = 10$ cells from three independent experiments, $P = 0.0102$. In **n**, $n = 10$ cells from three independent experiments, $P = 0.0022$. β-arrestin mutations are described in detail in Extended Data Table 1. Scatter plots show overlay of mean and s.e.m. Scale bars, 5 μm. $^{**}P < 0.01$; $^{***}P < 0.001$.

(sptPALM[37], Fig. 5a–d, Extended Data Fig. 8a–e, Supplementary Videos 1–4). β-arrestin-2 diffusion profiles exhibited two major peaks when activated by the β2AR. A mobile fraction ($D > 10^{-2}\ \mu\text{m}^2\ \text{s}^{-1}$; $\log D > -2$) overlapped the major diffusion peak of the β2AR, and also that of a lipid probe (PH-PLCδ1) that is known to diffuse freely in the plasma membrane[38] (Fig. 5e, Extended Data Fig. 8f–h). An immobile fraction ($D \le 10^{-2}\ \mu\text{m}^2\ \text{s}^{-1}$; $\log D \le -2$), consistent with the mobility of CCSs was also observed[39], and the density of immobile β-arrestin-2 molecules was significantly higher within a CCS mask compared to the rest of the plasma membrane. This was true irrespective of whether β-arrestin-2 trafficking was activated by the β1AR or β2AR (9.3 and 6.0-fold enrichment, respectively; $n = 11$ and 9 cells, respectively, from three independent experiments; $P = 0.0004$ and 0.0016, respectively; statistical significance calculated using a two-tailed unpaired *t*-test). Destabilizing clathrin and AP-2 interactions using the β-arrestin-2 CCS mutant shifted the distribution into the mobile peak, irrespective of whether β-arrestin-2 recruitment was activated by the β1AR or β2AR (Fig. 5f, g, Extended Data Fig. 8i–l, Supplementary Videos 5, 6), confirming that the immobile fraction of β-arrestin-2 molecules largely represents those bound to the CCS lattice. This suggests that disrupting CCS binding displaces β-arrestin towards an association with a diffusive partner on the plasma membrane. Consistent with this, β-arrestin-2 molecules present within CCSs were essentially immobile after activation by either the β1AR or β2AR ($\log D = -2.3$ and -2.7, respectively; Extended Data Fig. 8m–p). Trajectories of both β1AR and β2AR populated CCSs, but β2ARs were preferentially immobilized ($\log D = -2.3$) there, whereas β1ARs remained largely mobile ($\log D = -1.6$). Together, these results indicate that β-arrestin is similarly immobilized in CCSs irrespective of whether (β2AR) or not (β1AR) the activating GPCR is also immobilized there.

Stability of β-arrestin capture

Our results support a working model in which the ligand-activated GPCR acts catalytically to activate β-arrestin trafficking (Extended Data Fig. 8q). β-arrestin is subsequently captured and stabilized at the plasma membrane after dissociating from its activating GPCR through a series of non-GPCR interactions that ultimately produce β-arrestin accumulation at CCSs. For this discrete mechanism to be energetically feasible, these non-GPCR interactions must be relatively stable. Indeed, fluorescence recovery after photobleaching (FRAP) demonstrated that β-arrestin-2 accumulated in CCSs exchanges very slowly (half time of equilibrium, $t_{1/2} > 60$ s), whether (β2AR) or not (β1AR) the GPCR stimulating β-arrestin accumulation in CCSs co-accumulated (Fig. 5h–j, Extended Data Fig. 8r). Notably, this duration of β-arrestin association with the CCS is comparable to the lifetime of individual CCSs[18]. Thus CCSs have the capacity to act both as 'sinks' to stabilize β-arrestin at the plasma membrane after GPCR dissociation and as 'drivers' of the discrete trafficking mechanism through β-arrestin dissociation from the plasma membrane coupled to endocytic scission of CCSs[18].

Catalysis and scaffolding co-exist

We believe that distinct catalytic and scaffold-driven mechanisms of GPCR-regulated β-arrestin trafficking are likely to coexist in vivo, with the tendency of a particular GPCR to engage one mechanism relative to the other tuned by the affinity of its tail for β-arrestin (Extended Data Fig. 9a). The β1AR appears to be a relatively pure example of a GPCR that activates β-arrestin trafficking primarily through the catalytic mechanism and with very little co-trafficking. The β2AR–V2R chimaera favours co-trafficking in a tail-dependent scaffold complex, with trafficking of β-arrestin to CCSs not requiring phosphoinositide binding. The β2AR has a mixed trafficking behaviour, producing super-stoichiometric trafficking of β-arrestin to CCSs that is

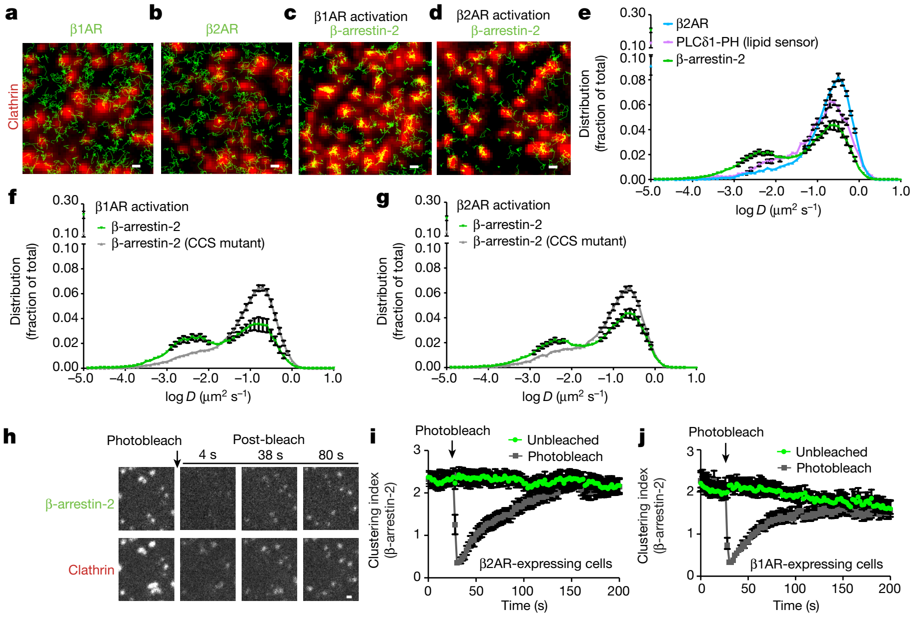


Fig. 5 | sptPALM analysis of GPCR and β-arrestin dynamics and stable β-arrestin binding at CCSs. a–d, Representative images (from at least three independent experiments) of PAmCherry–β1AR (**a**; green) or PAmCherry–β2AR (**b**; green) trajectories and PAmCherry–β-arrestin-2 trajectories (green) with β1AR expression (**c**) or PAmCherry–β-arrestin-2 (green) trajectories with β2AR expression (**d**) from sptPALM analysis overlaid with a clathrin marker (red) after treatment with 10 μM isoproterenol. **e,** False-positive-corrected diffusion coefficients (*D*) of PAmCherry–β2AR, β-arrestin-2–PAmCherry and PAmCherry–PLCδ1-PH in live cells after treatment with 10 μM isoproterenol ($n = 13$, 21 and 8 cells, respectively). β-arrestin-2–PAmCherry and PAmCherry–PLCδ1-PH were co-expressed individually with Flag–β2AR. PLCδ1-PH is the PH domain of phospholipase C δ1, which selectively binds to PtdIns(4,5)P_2. **f, g,** False-positive-corrected distribution of diffusion coefficients of wild-type and CCS mutant β-arrestin-2–PAmCherry when co-expressed with Flag–β1AR (**f**; $n = 13$ and 17 cells, respectively, from three independent experiments; statistical significance of the immobile fractions was calculated using a two-tailed unpaired *t*-test, $P < 0.0001$) or Flag–β2AR (**g**; $n = 21$ and 10 cells, respectively, from three independent experiments; statistical significance of the immobile fractions was calculated using a two-tailed unpaired *t*-test, $P = 0.002$). β-arrestin-2–PAmCherry diffusion coefficient profiles when activated by the β2AR are replotted from **e**. **h,** COS-1 cells co-expressing Flag–β2AR, β-arrestin-2–GFP (green), and clathrin light chain–DsRed (red) were treated with 10 μM isoproterenol for 3 min before photobleaching of β-arrestin-2–GFP. Shown are representative images (from three independent experiments) of the photobleached area. β-arrestin-2 clustering index over the course of the photobleaching experiment in cells co-expressing activated Flag–β1AR (**i**; $n = 12$ cells from three independent experiments) or Flag-β2AR (**j**; $n = 15$ and 13 cells for unbleached and photobleached conditions, respectively, from three independent experiments). Data are mean ± s.e.m. Scale bars, 0.5 μm.

dependent on phosphoinositide binding, yet also co-trafficking to CCSs through interaction of the phosphorylated receptor tail with β-arrestin. Manipulations that enhance GPCR tail binding to β-arrestin are able to further increase β2AR co-trafficking to CCSs and also cause β1AR to co-traffic (Extended Data Fig. 9b–e), emphasizing the tunability of this system.

Discussion

These results reveal a new framework of cellular β-arrestin activation but raise many additional questions. One is whether catalytic activation of β-arrestin produces a similar or different conformational state relative to β-arrestin that traffics in a tail-dependent GPCR–β-arrestin scaffold complex. Tail-dependent GPCR–β-arrestin complex formation is thought to promote an open β-arrestin conformation defined by displacement of the C terminus of β-arrestin. Determinants of binding to clathrin and AP-2 that are required for β-arrestin accumulation in CCSs are located in the C terminus; therefore, β-arrestin that accumulates in CCSs via the catalytic activation mechanism is likely to be in an open conformation. However, it is not known whether other conformational features of β-arrestin are similar or different between the mechanisms. The accompanying Letter by Latorraca et al.[40] provides insight into this question by demonstrating through molecular dynamics analysis not only that the GPCR core can promote conformational activation of arrestins, but also that arrestins can frequent conformations similar to those found in the GPCR tail-bound complex once the arrestin C terminus is dissociated, even without the GPCR being bound. Another important goal is to more fully resolve individual steps in the catalytic activation mechanism and determine their kinetics. For example, β-arrestin trafficking could involve additional steps such as transactivation or binding to the phosphorylated tail of another GPCR. Much also remains unknown about the functional consequences of GPCR-catalysed β-arrestin trafficking. β1ARs are known to promote β-arrestin-dependent ERK signalling by activating β-arrestin trafficking to CCSs without the receptor[18]; we verified this using a CRISPR knockout approach[41] (Extended Data Fig. 9f, g). The β2AR is unlike the β1AR in that it has a tendency to co-traffic to CCSs with β-arrestin; and β2AR did not produce a strong β-arrestin-dependent component of ERK activation using this assay, when examined in parallel and with matched receptor expression levels (Extended Data Fig. 9h, i). Therefore we anticipate that the presently delineated catalytic mechanism of cellular β-arrestin activation, by enabling β-arrestin to traffic to CCSs separately from its upstream activating GPCR, is likely to have widespread physiological significance.

Online content

Any Methods, including any statements of data availability and Nature Research reporting summaries, along with any additional references and Source Data files, are available in the online version of the paper at https://doi.org/10.1038/s41586-018-0079-1.

Received: 24 August 2017; Accepted: 20 March 2018;

1. Rosenbaum, D. M., Rasmussen, S. G. F. & Kobilka, B. K. The structure and function of G-protein-coupled receptors. *Nature* **459**, 356–363 (2009).
2. Lohse, M. J., Benovic, J. L., Codina, J., Caron, M. G. & Lefkowitz, R. J. beta-Arrestin: a protein that regulates beta-adrenergic receptor function. *Science* **248**, 1547–1550 (1990).

3. Kang, D. S., Tian, X. & Benovic, J. L. Role of β-arrestins and arrestin domain-containing proteins in G protein-coupled receptor trafficking. *Curr. Opin. Cell Biol.* **27**, 63–71 (2014).
4. Pierce, K. L., Premont, R. T. & Lefkowitz, R. J. Seven-transmembrane receptors. *Nat. Rev. Mol. Cell Biol.* **3**, 639–650 (2002).
5. Gurevich, E. V. & Gurevich, V. V. Arrestins: ubiquitous regulators of cellular signaling pathways. *Genome Biol.* **7**, 236 (2006).
6. Lohse, M. J. & Hoffmann, C. *Arrestins–Pharmacology and Therapeutic Potential* (Springer, Berlin, Heidelberg, 2014).
7. Shukla, A. K., Xiao, K. & Lefkowitz, R. J. Emerging paradigms of β-arrestin-dependent seven transmembrane receptor signaling. *Trends Biochem. Sci.* **36**, 457–469 (2011).
8. Gurevich, V. V. & Gurevich, E. V. The structural basis of arrestin-mediated regulation of G-protein-coupled receptors. *Pharmacol. Ther.* **110**, 465–502 (2006).
9. Nuber, S. et al. β-Arrestin biosensors reveal a rapid, receptor-dependent activation/deactivation cycle. *Nature* **531**, 661–664 (2016).
10. Cahill, T. J. III et al. Distinct conformations of GPCR-β-arrestin complexes mediate desensitization, signaling, and endocytosis. *Proc. Natl Acad. Sci. USA* **114**, 2562–2567 (2017).
11. Lee, M.-H. et al. The conformational signature of β-arrestin2 predicts its trafficking and signalling functions. *Nature* **531**, 665–668 (2016).
12. Kumari, P. et al. Functional competence of a partially engaged GPCR–β-arrestin complex. *Nat. Commun.* **7**, 13416 (2016).
13. Kumari, P. et al. Core engagement with β-arrestin is dispensable for agonist-induced vasopressin receptor endocytosis and ERK activation. *Mol. Biol. Cell* **28**, 1003–1010 (2017).
14. Branco, A. F. et al. Isoproterenol cytotoxicity is dependent on the differentiation state of the cardiomyoblast H9c2 cell line. *Cardiovasc. Toxicol.* **11**, 191–203 (2011).
15. Puthenveedu, M. A. & von Zastrow, M. Cargo regulates clathrin-coated pit dynamics. *Cell* **127**, 113–124 (2006).
16. Santini, F., Gaidarov, I. & Keen, J. H. G protein-coupled receptor/arrestin3 modulation of the endocytic machinery. *J. Cell Biol.* **156**, 665–676 (2002).
17. Barak, L. S., Ferguson, S. S., Zhang, J. & Caron, M. G. A β-arrestin/green fluorescent protein biosensor for detecting G protein-coupled receptor activation. *J. Biol. Chem.* **272**, 27497–27500 (1997).
18. Eichel, K., Jullié, D. & von Zastrow, M. β-Arrestin drives MAP kinase signalling from clathrin-coated structures after GPCR dissociation. *Nat. Cell Biol.* **18**, 303–310 (2016).
19. Mondin, M. et al. Neurexin-neuroligin adhesions capture surface-diffusing AMPA receptors through PSD-95 scaffolds. *J. Neurosci.* **31**, 13500–13515 (2011).
20. Shukla, A. K. et al. Structure of active β-arrestin-1 bound to a G-protein-coupled receptor phosphopeptide. *Nature* **497**, 137–141 (2013).
21. Shukla, A. K. et al. Visualization of arrestin recruitment by a G-protein-coupled receptor. *Nature* **512**, 218–222 (2014).
22. Xiao, K., Shenoy, S. K., Nobles, K. & Lefkowitz, R. J. Activation-dependent conformational changes in β-arrestin 2. *J. Biol. Chem.* **279**, 55744–55753 (2004).
23. Nobles, K. N., Guan, Z., Xiao, K., Oas, T. G. & Lefkowitz, R. J. The active conformation of β-arrestin1: direct evidence for the phosphate sensor in the N-domain and conformational differences in the active states of β-arrestins1 and -2. *J. Biol. Chem.* **282**, 21370–21381 (2007).
24. McCorvy, J. D. et al. Structure-inspired design of β-arrestin-biased ligands for aminergic GPCRs. *Nat. Chem. Biol.* **14**, 126–134 (2018).
25. Gimenez, L. E., Babilon, S., Wanka, L., Beck-Sickinger, A. G. & Gurevich, V. V. Mutations in arrestin-3 differentially affect binding to neuropeptide Y receptor subtypes. *Cell. Signal.* **26**, 1523–1531 (2014).
26. Violin, J. D., Ren, X.-R. & Lefkowitz, R. J. G-protein-coupled receptor kinase specificity for β-arrestin recruitment to the β_2-adrenergic receptor revealed by fluorescence resonance energy transfer. *J. Biol. Chem.* **281**, 20577–20588 (2006).
27. Peterson, S. M. et al. Elucidation of G-protein and β-arrestin functional selectivity at the dopamine D2 receptor. *Proc. Natl Acad. Sci. USA* **112**, 7097–7102 (2015).
28. Gurevich, V. V. The selectivity of visual arrestin for light-activated phosphorhodopsin is controlled by multiple nonredundant mechanisms. *J. Biol. Chem.* **273**, 15501–15506 (1998).
29. Scheerer, P. & Sommer, M. E. Structural mechanism of arrestin activation. *Curr. Opin. Struct. Biol.* **45**, 160–169 (2017).
30. Chen, Q. et al. Structural basis of arrestin-3 activation and signaling. *Nat. Commun.* **8**, 1427 (2017).
31. Gaidarov, I., Krupnick, J. G., Falck, J. R., Benovic, J. L. & Keen, J. H. Arrestin function in G protein-coupled receptor endocytosis requires phosphoinositide binding. *EMBO J.* **18**, 871–881 (1999).
32. Lally, C. C. M., Bauer, B., Selent, J. & Sommer, M. E. C-edge loops of arrestin function as a membrane anchor. *Nat. Commun.* **8**, 14258 (2017).
33. Goodman, O. B. Jr et al. β-arrestin acts as a clathrin adaptor in endocytosis of the β_2-adrenergic receptor. *Nature* **383**, 447–450 (1996).
34. Laporte, S. A., Oakley, R. H., Holt, J. A., Barak, L. S. & Caron, M. G. The interaction of β-arrestin with the AP-2 adaptor is required for the clustering of β_2-adrenergic receptor into clathrin-coated pits. *J. Biol. Chem.* **275**, 23120–23126 (2000).
35. Oakley, R. H., Laporte, S. A., Holt, J. A., Barak, L. S. & Caron, M. G. Association of β-arrestin with G protein-coupled receptors during clathrin-mediated endocytosis dictates the profile of receptor resensitization. *J. Biol. Chem.* **274**, 32248–32257 (1999).
36. Santos, M. de, S., Naal, R. M. Z. G., Baird, B. & Holowka, D. Inhibitors of PI(4,5)P2 synthesis reveal dynamic regulation of IgE receptor signaling by phosphoinositides in RBL mast cells. *Mol. Pharmacol.* **83**, 793–804 (2013).
37. Manley, S. et al. High-density mapping of single-molecule trajectories with photoactivated localization microscopy. *Nat. Methods* **5**, 155–157 (2008).
38. Hammond, G. R. V., Sim, Y., Lagnado, L. & Irvine, R. F. Reversible binding and rapid diffusion of proteins in complex with inositol lipids serves to coordinate free movement with spatial information. *J. Cell Biol.* **184**, 297–308 (2009).
39. Liu, A. P., Loerke, D., Schmid, S. L. & Danuser, G. Global and local regulation of clathrin-coated pit dynamics detected on patterned substrates. *Biophys. J.* **97**, 1038–1047 (2009).
40. Latorraca, N. R. et al. Molecular mechanism of GPCR-mediated arrestin activation. *Nature* https://doi.org/10.1038/s41586-018-0077-3 (2018).
41. O'Hayre, M. et al. Genetic evidence that β-arrestins are dispensable for the initiation of β_2-adrenergic receptor signaling to ERK. *Sci. Signal.* **10**, eaal3395 (2017).

Acknowledgements We thank J. Benovic, V. Gurevich, A. Inoue and S. Gutkind for sharing reagents and discussions; T. Balla, J. Fraser, B. Kobilka, R. Lefkowitz, M. Lohse, A. Manglik, B. Shoichet, M. Sommer, R. Sunahara and B. Lobingier, K. Varandas and other von Zastrow laboratory members for discussions; and W. Huynh and R. Vale for contributions to protein purification. All live-cell imaging experiments were performed in the Nikon Imaging Center at UCSF. These studies were supported by grants from the U.S. National Institutes of Health (to R.O.D. and M.v.Z.). J.-B.S. received support from Aquitaine grant Dynascreen, LabEx BRAIN and the IdEx Bordeaux. K.E. and N.R.L. are recipients of National Science Foundation Graduate Research Fellowships. M.M. is a recipient of an American Heart Association Postdoctoral Fellowship.

Reviewer information *Nature* thanks D. Calebiro, C. Hoffmann and B. Roth for their contribution to the peer review of this work.

Author contributions K.E., D.J., B.B.-R., N.R.L., R.O.D. and M.v.Z designed the research. D.J. performed and analysed single-molecule imaging experiments. B.B.-R. performed and analysed FRAP imaging experiments. K.E. performed and analysed all other imaging and biochemical experiments. N.R.L. performed and analysed molecular dynamics simulations and provided a structural interpretation of experimental data, with guidance from R.O.D. M.M. provided important advice and assistance with in vitro studies. J.-B.S provided software and advice for single molecule analysis. K.E. and M.v.Z. wrote the paper with input from all authors.

Competing interests The authors declare no competing interests.

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0079-1.

Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0079-1.

Reprints and permissions information is available at http://www.nature.com/reprints.

Correspondence and requests for materials should be addressed to M.v.Z.

METHODS

Cell culture, expression constructs, and transfections. HEK 293, COS-1, and H9c2 cells (ATCC authenticated lines CRL-1573, CRL-1650 and CRL 1446, respectively) were cultured in complete growth Dulbecco's modified Eagle's medium (DMEM, Gibco) supplemented with 10% fetal bovine serum (UCSF Cell Culture Facility). Cell line cultures were free of mycoplasma contamination. Transfections were carried out using Lipofectamine 2000 for cDNA according to the manufacturer's protocol. Cells were transfected 48 h before experiments.

N-terminally Flag-tagged versions of the human β1AR, β2AR, μ-opioid receptor (MOR) and κ-opioid receptor (KOR) were previously described[42,43]. Super ecliptic pHluorin–β2AR was previously described[44]. β1AR and β2AR tagged N-terminally with photoactivatable mCherry (PAmCherry–β1AR, PAmCherry–β2AR) was generated using PCR and homology-directed ligation (In-Fusion HD Cloning kit, Clontech). DRD2 was a gift from D. Grandy (Oregon Health & Science University). DRD2 (G prot) was prepared as previously described[27] by inserting a gBlock gene fragment (IDT) containing the desired mutations with restriction sites BamHI and BstEII (NEB) into the wild-type DRD2 plasmid. The β2AR–V2R chimaera[45] was a gift from M. Caron. β1AR–β2AR C tail, β1AR–V2R C tail, β2AR–β1AR C tail were generated using PCR and homology-directed ligation (In-Fusion HD Cloning kit, Clontech). N-terminally Flag-tagged β1AR(415T) was generated from Flag–β1AR using site-directed mutagenesis (Phusion Site-Directed Mutagenesis Kit, Thermo Scientific) to create a deletion after residue 415. N-terminally Flag-tagged β2AR(341T) and Flag-tagged β2AR(365T), which were previously described[46,47], were prepared by PCR site-directed mutagenesis (Phusion Site-Directed Mutagenesis Kit, Thermo Scientific). β2AR with an N-terminal Flag tag and a C-terminal GFP tag was generated by inserting a gBlock (IDT) containing a C-terminal fragment of β2AR, a 12 base pair linker and GFP using EcoRV and PshAI (NEB).

β-arrestin-2–GFP, β-arrestin-2–mApple and β-arrestin-2–PAmCherry were previously described[17,18]. β-arrestin-1–mVenus was a gift from R. Sunahara (University of California, San Diego). Specific β-arrestin-2–GFP mutations are described in detail in Extended Data Table 1. All β-arrestin-2–GFP finger-loop-proximal mutants were created by inserting a gBlock (IDT) containing the desired mutations into the β-arrestin-2–GFP wild-type plasmid using restriction sites HindIII and BbvCI (NEB). β-arrestin-2–GFP(E314K) and β-arrestin-2–GFP (K77E/E314K) were created using site-directed mutagenesis (Phusion Site-Directed Mutagenesis Kit, Thermo Scientific) from the wild-type β-arrestin-2–GFP or β-arrestin-2–GFP (K77E) construct, respectively. β-arrestin-2–GFP lipid-binding mutant was previously described[31] and was created by inserting a gBlock (IDT) containing the desired mutations into the β-arrestin-2–GFP wild-type plasmid using restriction sites BbvCI and AhdI (NEB). β-arrestin-2–GFP CCS mutant and β-arrestin-2–PAmCherry CCS mutant were created by inserting a gBlock (IDT) containing the desired mutations into the β-arrestin-2–GFP and β-arrestin-2–PAmCherry plasmids, respectively, using restriction sites BlpI and ApaI (NEB). β-arrestin-2–GFP lipid and CCS mutant and β-arrestin-2–GFP lipid and finger-loop-proximal mutant constructs were created by inserting a gBlock (IDT) containing the desired mutations into the β-arrestin-2–GFP CCS mutant plasmid and β-arrestin-2–GFP finger-loop-proximal mutant construct, respectively, using restriction sites BbvCI and AhdI (NEB). β-arrestin-2–GFP(L191G/F192G) was created using site-directed mutagenesis (Phusion Site-Directed Mutagenesis Kit, Thermo Scientific) from the wild-type β-arrestin-2–GFP. β-arrestin-1–mVenus finger-loop-proximal mutant was cloned by inserting a gBlock (IDT) containing the desired mutation into the β-arrestin-1–mVenus wild-type plasmid using restriction sites BamHI and SphI (NEB). β-arrestin-2–GFP KNC mutant, which has been previously described and shown to be defective in binding to GPCRs[25], was subcloned into β-arrestin-2–GFP from a β-arrestin-2 expression plasmid (a gift from V. Gurevich). NAV3β–arrestin-1(1–393) in pGEX4T was generated from a previously described construct[23] and contains an N-terminal GST tag, 3C cleavage site and AVI tag. NAV3β–arrestin-1(1–393) lipid-binding mutant was created by inserting a gBlock (IDT) containing the desired mutations into the NAV3β–arrestin-1(1–393) using restriction sites EcoRI and NcoI (NEB).

Clathrin–dsRed, clathrin–GFP and clathrin–TagBFP were previously described[15,18,48]. GRK2–tagBFP was created by PCR amplifying GRK2, which was a gift from J. Benovic (Thomas Jefferson University), and subcloning into the pTagBFP vector (Evrogen) using NheI and SacII (NEB). PAmCherry–PLCδ1-PH was subcloned from GFP–PLCδ1-PH[49], a gift from T. Meyer (addgene plasmid #21179), using BspEI and EcoRI (NEB).

Live cell TIRF microscopy imaging. TIRF microscopy was performed at 37 °C using a Nikon Ti-E inverted microscope equipped for through-the-objective TIRF microscopy and outfitted with a temperature-, humidity- and CO2-controlled chamber (Okolab). Images were obtained with an Apo TIRF 100 × , 1.49 numerical aperture objective (Nikon) with solid-state 405, 488, 561 and 647 nm lasers (Keysight Technologies). An Andor iXon DU897 EMCCD camera controlled by NIS-Elements 4.1 software was used to acquire image sequences every 2 s for 10 min. Unless indicated otherwise, live-cell microscopy assays were performed using HEK 293 cells. Cells were transfected as indicated according to the manufacturer's protocol 48 h before imaging and then plated on poly-L-lysine (0.0001%, Sigma) coated 35-mm glass-bottomed culture dishes (MatTek Corporation) 24 h before imaging. Cells were labelled with monoclonal Flag antibody (M1) (1:1000, Sigma F-3040) conjugated to Alexa Fluor 647 dye (Life Technologies) for 10 min at 37 °C before imaging, washed, and imaged live in DMEM without phenol red (UCSF Cell Culture Facility) supplemented with 30 mM HEPES, pH 7.4 (UCSF Cell Culture Facility). Cells were treated with bath application of the indicated agonist at time 0 s for experiments shown as timecourses. At least three independent experiments were performed for all live-cell TIRF microscopy imaging.

TIRF microscopy image analysis. Quantitative image analysis was performed on unprocessed images using ImageJ and Fiji software[50,51]. To quantify change in β-arrestin fluorescence over time in TIRF microscopy images, which was reported as plasma membrane recruitment, fluorescence values were measured over the entire stack in a region of interest (ROI) corresponding to the cell. Fluorescence values of the ROI were normalized to initial fluorescence values before agonist addition. Minimal bleed-through and photobleaching was verified using single-labelled and untreated samples, respectively. Linescan analysis of receptor, β-arrestin, or clathrin fluorescence from the shown line were carried out using the Fiji plot profile function to measure pixel values from this line. For calculations of fluorescence enrichment into CCSs, a mask of CCSs was generated using a thresholded average image of the clathrin channel. Enrichment at CCSs for receptor and arrestin was measured as the difference between the average fluorescence in the mask and average fluorescence outside of the thresholded structures. Clustering index was determined using the skew statistical measurement applied to fluorescence intensity values of β-arrestin–GFP pixels in a ROI corresponding to the cell.

Fluorescence recovery after photobleaching. FRAP was performed at 37 °C using a Nikon Ti inverted spinning disk confocal microscope (Yokogawa CSU-W1) equipped with a temperature-, humidity- and CO2-controlled chamber (Okolab). Images were obtained with a Plan Apo VC 100 × , 1.4 NA objective (Nikon) with 488-, 561- and 640-nm solid-state lasers (Keysight Technologies). An Andor Zyla 4.2 sCMOS camera controlled by MicroManager 2.0 software was used to acquire image sequences. A Rapp Optoelectronic UGA-40 photobleaching system was used to photobleach β-arrestin-2–GFP on a small area of the plasma membrane with a 473-nm laser (Vortran). Cells were imaged every 2 s for 10 min to monitor fluorescence recovery after photobleaching. All quantitative image analysis was performed on unprocessed images using ImageJ software. The clustering index was calculated for the photobleached and unbleached areas of identical size in the same cell. Photobleaching experiments were performed at least three independent times.

Quantitative live-cell sptPALM. Cells were transfected as indicated according to the manufacturer's protocol 48 h before imaging and then plated on poly-L-lysine (0.0001%, Sigma) coated 35-mm glass-bottomed culture dishes (MatTek Corporation) 24 h before imaging. Results were obtained with at least three independent experiments, except for the analysis of PAmCherry–β1AR diffusion coefficients profile. For experiments to investigate the diffusion coefficient profile of PAmCherry-tagged receptors, lipid sensor, β-arrestin-2 or control cells without PAmCherry protein expression, cells were surface-labelled with M1–Alexa Fluor 488 and experiments were performed blind regarding the transfection condition. For experiments to localize receptors or β-arrestin behaviour relative to CCSs, cells were incubated for 10 min with 100-nm TetraSpeck microspheres (ThermoFisher) and imaging was performed blind regarding transfection condition. sptPALM experiments were performed at 37 °C using a Nikon Ti-E inverted microscope equipped with TIRF illumination and outfitted with a temperature-, humidity- and CO2-controlled chamber (Okolab). Cells were imaged in a solution containing 135 mM NaCl, 5 mM KCl, 0.4 mM $MgCl_2$, 1.8 mM $CaCl_2$, 20 mM HEPES and 5 mM D-glucose, adjusted to pH 7.4, 1 min after addition of 10 μM isoproterenol. Images were acquired with a PL-Apo TIRF 100 × , 1.49 NA objective (Nikon) with solid-state 405-, 488-, 561- and 647-nm lasers (Keysight Technologies) as light sources. An Andor iXon DU897 EMCCD camera controlled by NIS-Elements 4.1 software was used to acquire stacks of 5,000 images with continuous activation by 405-nm light and imaging by 561-nm laser. Single-molecule image sequences were acquired in streaming mode at 45 Hz in a 256 × 256-pixel area (40 × 40-μm field of view). For colocalization experiments, a clathrin–GFP image was aquired at the end of the sptPALM image series.

Single-molecule fluorescent spots were localized and tracked over time using a combination of wavelet-based segmentation and simulated-annealing tracking algorithms as previously described[52,53]. The software package used to derive quantitative data on protein localization and dynamics is custom written, operating as a plug-in running within the MetaMorph software (Molecular Devices) environment. The mean square displacement (MSD) and diffusion coefficient (D) were calculated for every trajectory as follows. For every trajectory of N data points (coordinates $x(t)$, $y(t)$ at times $t = 0$ to $N\Delta t$ with $\Delta t = 22$ ms), the mean square displacement (MSD) for time intervals $\tau = n \times \Delta t$ is calculated using the formula

$$\sum_{i=1}^{N-n} \frac{[x((i+n)\times\Delta t) - x(i\times\Delta t)]^2 + [y((i+n)\times\Delta t) - y(i\times\Delta t)]^2}{N-n}$$

D was then extracted by linear fit on the first four time points of the MSD curves using the formulae $\mathrm{MSD}(\tau) = \langle r^2 \rangle (\tau) + \alpha = 4D\tau + \alpha$, in which α is the ordinate at the origin of the linear fit due to the precision accuracy in the localization process. The bin at $10^{-5}\,\mu\mathrm{m}^2\,\mathrm{s}^{-1}$ represents trajectories for which calculation of $D \leq 10^{-5}\,\mu\mathrm{m}^2\,\mathrm{s}^{-1}$.

To define the diffusion profile across the different conditions, we included only trajectories longer than six points. To take into account false detections due to non-specific single-molecule signals, we computed an average histogram of diffusion coefficients from trajectories obtained by imaging cells without PAmCherry expression, using the same parameters described previously. The variability of the number of localizations and diffusion profiles of non-specific single-molecule signal was very low. For each cell exhibiting at least five times more trajectories than the average false-detection count, the histogram of false detections was subtracted from the distribution of diffusion coefficients. Histograms of diffusion coefficients were computed by normalizing the number of trajectories for each bin by the total number of trajectories after false-detection subtraction. We considered trajectories with $D \leq 10^{-2}\,\mu\mathrm{m}^2\,\mathrm{s}^{-1}$ as immobile and used this criterion for evaluation of statistical significance using a two-tailed *t*-test.

For analysis of localization of trajectories relative to CCPs, single-particle tracking movies as well as diffraction limited clathrin-GFP images were aligned with subpixel accuracy using a Gaussian fitting on the Tetraspeck microsphere signal as fiduciary marker. Trajectories of more than eight points were used to generate the super-resolution images with a pixel size of 19.6 nm (zoom 8 × compared to the acquired data) and compute the diffusion coefficients as described above. A series of super-resolution images was generated displaying reconstructed individual trajectories, localization density of mobile molecules ($D \geq 10^{-2}\,\mu\mathrm{m}^2\,\mathrm{s}^{-1}$), immobile molecules ($D \leq 10^{-2}\,\mu\mathrm{m}^2\,\mathrm{s}^{-1}$), or diffusion coefficient maps. Diffusion coefficient maps were computed by averaging in each pixel the diffusion coefficients from all the trajectories detected in the corresponding pixel. Images of CCSs were zoomed 8 × before alignment and overlaid with the super-resolution images. A binary mask of the CCSs was generated using ImageJ (NIH) by thresholding and eroding by two pixels the 8 × zoomed image of the clathrin mask after bandpass filtering in the Fourier domain. This mask was used in combination with the super-resolution images to calculate the average diffusion coefficient as well as the density of localization with respect to the CCSs.

ERK1/2 activation assays. Western blot analysis was used to measure activation of ERK1/2. Previously described parental HEK 293 and β-arrestin CRISPR knockout HEK 293 cell lines[41] were transfected with empty vector control, Flag–β1AR or Flag–β2AR expression constructs. Cells were serum starved for 18 h before the assay, incubated with 10 μM isoproterenol for the indicated times at 37 °C, and then washed on ice with ice-cold PBS. Cells were directly lysed in sample buffer (NuPAGE LDS Sample Buffer (Life Technologies), 100 mM dithiothreitol), sonicated three times for 10 s, boiled, separated by SDS–PAGE (Life Technologies) and transferred to a nitrocellulose membrane that was blocked with TBS Odyssey blocking buffer (LI-COR) for one hour at room temperature and then incubated overnight at 4 °C with a mouse ERK1/2 primary antibody (1:2,000, Cell Signaling 4696) and a rabbit phosphorylated-ERK1/2 primary antibody (1:2,000, Cell Signaling 4370). Membranes were washed three times for 5 min in TBS-Tween (0.1% v/v) and incubated with a IRDye 680-labelled anti-rabbit secondary antibody (1:5,000, LI-COR 926-68073) and an IRDye 800-labelled anti-mouse secondary antibody (1:5,000, LI-COR 926-32212) for one hour at room temperature. Membranes were washed three times for 5 min in TBS-Tween (0.1% v/v), imaged using an Odyssey Infrared Imaging System (LI-COR) in the linear range and quantified by measuring band intensity, background subtracting, and normalizing the phosphorylated ERK1/2 band intensity to the total ERK1/2 band intensity. Data are shown as fraction of the maximum response observed across all conditions in each experiment. Five independent experiments were performed for each condition.

Co-immunoprecipitation. Cells expressing indicated constructs were grown to confluency in 10-cm dishes, and 48 h after transfection, cells were washed twice with reaction buffer (PBS, 30 mM HEPES, pH 7.4) and then crosslinked with 2 mM DSP (Thermo Scientific) for 30 min at room temperature with gentle agitation. The crosslinking reaction was stopped by addition of 20 mM Tris, pH 7.5 for 15 min. Cells were collected, pelleted by centrifugation, lysed on ice for 10 min in 500 μl ice-cold lysis buffer (0.2% Triton X-100, 150 mM NaCl, 25 mM KCl, 10 mM Tris pH 7.4, and 1 mM EDTA supplemented with a standard protease inhibitor mixture (Roche Applied Science)) and then cleared by centrifugation (14,000*g* for 15 min at 4 °C).

Samples were incubated overnight at 4 °C with anti-GFP antibody covalently linked to sepharose beads (1:3, Abcam ab69314), washed with lysis buffer three times, and incubated with SDS sample buffer (Invitrogen) supplemented with 100 mM dithiothreitol to elute proteins. Samples were then separated by SDS–PAGE (Life Technologies) and transferred to a nitrocellulose membrane that was blocked with TBS Odyssey blocking buffer (LI-COR) for one hour at room temperature and then incubated overnight at 4 °C with a mouse β-adaptin primary antibody (1:250, BD Biosciences 610382) and a rabbit GFP primary antibody (1:500, Thermo Fisher Scientific A-11122). Membranes were washed three times for 5 min in TBS-Tween (0.1% v/v) and incubated with an IRDye 680-labelled anti-rabbit secondary antibody (1:5,000, LI-COR 926-68073) and an IRDye 800-labelled anti-mouse secondary antibody (1:5,000, LI-COR 926-32212) for one hour at room temperature. Membranes were washed three times for 5 min in TBS-Tween (0.1% v/v), imaged using an Odyssey Infrared Imaging System (LI-COR) in the linear range and quantified by measuring band intensity, background subtracting, and normalizing the AP-2-immunoprecipitated band intensity to the GFP-bead band intensity. Three independent experiments were performed.

Purification of β-arrestin-1. N-terminally GST-tagged rat β-arrestin-1 (amino acids 1–393) constructs in the pGEX4T vector were transformed into BL21-CodonPlus(DE3)-RIPL cells (Agilent). Cultures were grown at 37 °C to an absorbance ($A_{600\,\mathrm{nm}}$) of 0.6 in Terrific broth and then equilibrated to 16 °C. GST-β-arrestin-1 expression was induced with 0.1 mM isopropyl 1-thio-β-D-galactopyranoside overnight at this temperature, and cells were collected by centrifugation at 6,000*g*. Pellets were resuspended with cold lysis buffer (50 mM HEPES pH 8.2, 150 mM NaCl, 2 mM DTT, protease inhibitors, 1 mM EDTA). Cells were lysed by passage through an Emulsiflex press (Avestin) and cleared by centrifugation at 40,000*g* for 50 min. The clarified supernatant was incubated with GST-4B resin (GE Healthcare) for 2 h at 4 °C, and then washed two times with five column volumes of wash buffer (50 mM HEPES pH7.4, 150 mM NaCl, 2 mM DTT). The GST resin with bound GST–β-arrestin-1 was resuspended in two column volumes of wash buffer, and the GST fusion protein was cleaved with 0.1 mg of 3C protease per ml of GST resin overnight at 4 °C. The supernatant and first wash fraction were pooled and centrifuged at 40,000*g* for 30 min. The cleared supernatant was filtered with a 0.45-μm filter, concentrated to 500 μl and gel-filtered using a Superose 6 10/300 gl column (GE Life Science). β-Arrestin-1 was eluted in gel filtration buffer (30 mM HEPES, 150 mM NaCl, 2 mM $MgCl_2$, 5% glycerol). Fractions were analysed by SDS–PAGE and fractions containing β-arrestin-1 were pooled and concentrated, flash-frozen in liquid N_2, and stored at −80 °C.

Lipid-bead binding. Equal amounts of purified wild-type or lipid-binding mutant β-arrestin-1 protein were incubated with PtdIns(4,5)P_2-coated beads (Echelon) for two hours at room temperature in wash/binding buffer (10 mM HEPES, pH 7.4, 0.25% NP-40, 150 mM NaCl). Samples were then washed three times with wash/binding buffer. To elute proteins, equal volumes of 2 × Laemmli sample buffer were added and samples were incubated at 95 °C for 5 min. Samples were then separated by SDS–PAGE (Life Technologies) and transferred to a nitrocellulose membrane that was blocked with TBS Odyssey blocking buffer (LI-COR) for one hour at room temperature and then incubated for one hour at room temperature with a mouse anti-β-arrestin primary antibody (1:500, Santa Cruz Biotechnology sc-13140). Membranes were washed three times for 5 min in TBS-Tween (0.1% v/v) and incubated with an IRDye 800-labelled anti-mouse secondary antibody (1:5,000, LI-COR 926-32212) for one hour at room temperature. Membranes were washed three times for 5 min in TBS-Tween (0.1% v/v), imaged using an Odyssey Infrared Imaging System (LI-COR) in the linear range, and quantified by measuring band intensity and subtracting background intensity. Four independent experiments were performed.

Molecular dynamics simulations. We analysed sets of molecular dynamics simulations for each of two conditions: (1) simulations initiated from the inactive-state β-arrestin-1 crystal structure (PDB ID: 1G4M, chain A), and (2) simulations initiated from the active-state β-arrestin-1 crystal structure bound to the phosphorylated C-tail of the V2 vasopressin receptor (PDB ID: 4JQI). For the latter condition, we removed the co-crystallized $F_{ab}30$ antibody fragment. We performed six simulations for each condition. For each simulation, initial atom velocities were assigned randomly and independently. These simulations correspond to simulations 40 to 51 in Supplementary Table 1 of the accompanying Letter[40], which also provides details regarding simulation setup and simulation protocols.

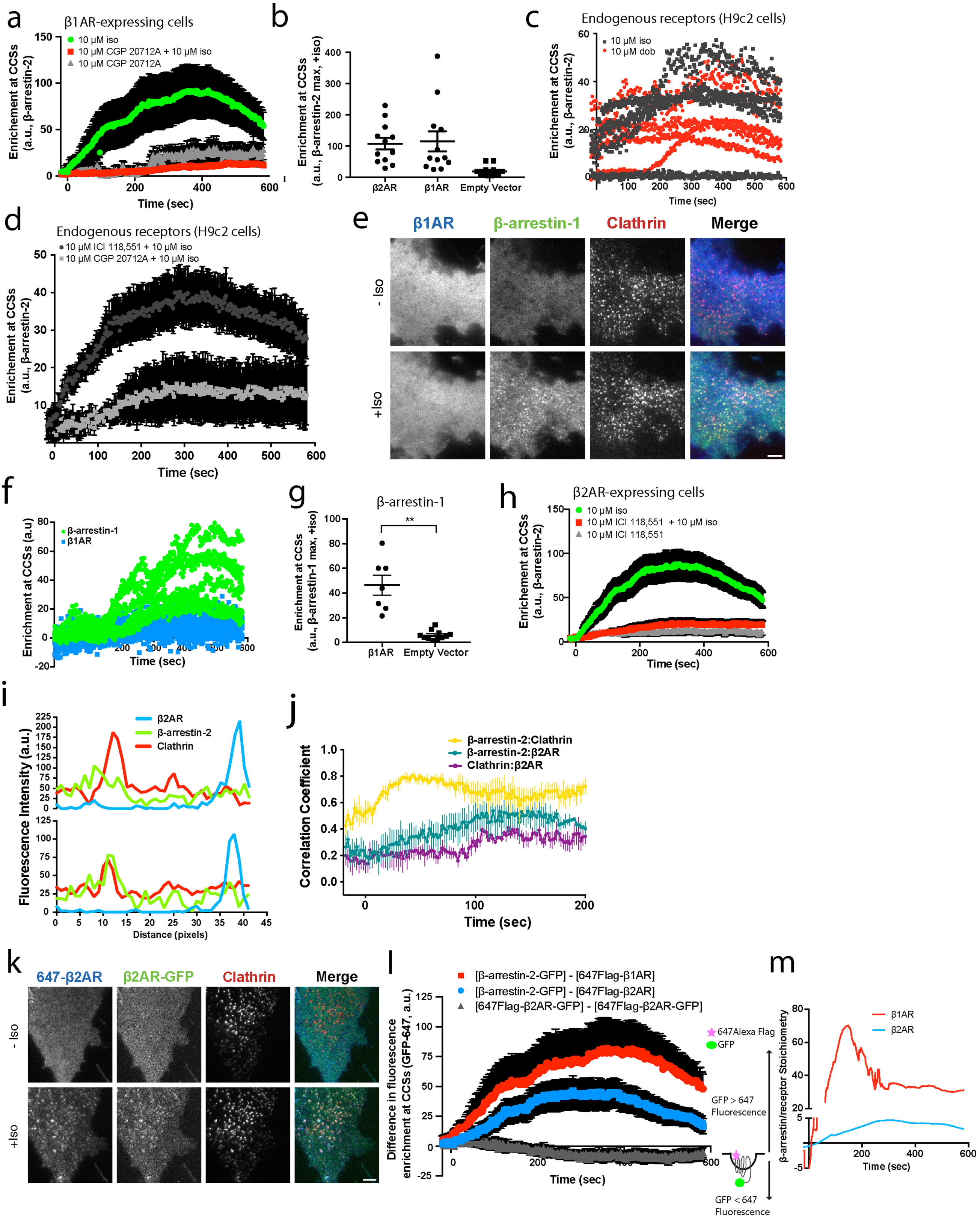
Simulations were visualized and analysed using Visual Molecular Dynamics (VMD)[54]. Simulations were inspected visually for interactions that formed in the inactive state but not the active state. We noticed that Glu313 occasionally formed a salt bridge with Lys77 in the inactive state, persisting for up to hundreds of nanoseconds. We quantified the frequency of salt bridge formation by calculating the minimum distance between polar heavy atoms of Lys77 and Glu313 across all simulations under each condition.

Statistical analysis. Quantitative data are expressed as the mean and error bars represent the standard error of the mean (s.e.m.) unless otherwise indicated. Scatter plots are overlaid with mean and s.e.m. Statistical significance between conditions was analysed using a two-tailed *t*-test or with Welch's correction for unequal variance and a one-way ANOVA ($\alpha = 0.05$) calculated using Prism 7.0 (GraphPad Software). $^*P < 0.05$; $^{**}P < 0.01$; $^{***}P < 0.001$ when compared with control or no-treatment conditions. All experiments showing representative data were repeated at least three independent times with similar results. Independent experiments represent independent biological replicates.

Reporting summary. Further information on experimental design is available in the Nature Research Reporting Summary linked to this paper.

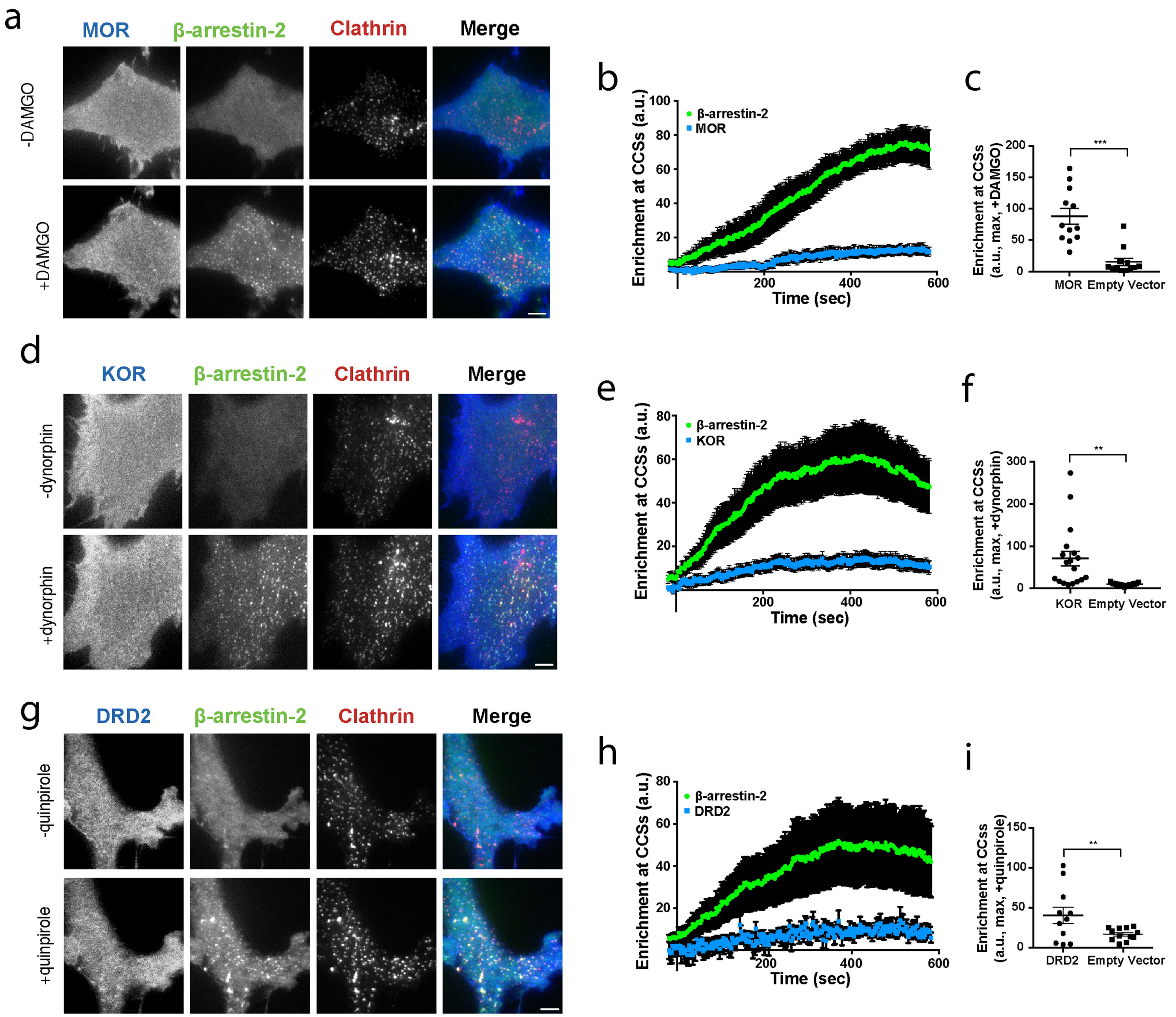
Data availability. The data that support the findings of this study are available from the authors upon reasonable request.

42. Cao, T. T., Deacon, H. W., Reczek, D., Bretscher, A. & von Zastrow, M. A kinase-regulated PDZ-domain interaction controls endocytic sorting of the β2-adrenergic receptor. *Nature* **401**, 286–290 (1999).
43. Temkin, P. et al. SNX27 mediates retromer tubule entry and endosome-to-plasma membrane trafficking of signalling receptors. *Nat. Cell Biol.* **13**, 715–721 (2011).
44. Yudowski, G. A., Puthenveedu, M. A. & von Zastrow, M. Distinct modes of regulated receptor insertion to the somatodendritic plasma membrane. *Nat. Neurosci.* **9**, 622–627 (2006).
45. Oakley, R. H., Laporte, S. A., Holt, J. A., Caron, M. G. & Barak, L. S. Differential affinities of visual arrestin, βarrestin1, and βarrestin2 for G protein-coupled receptors delineate two major classes of receptors. *J. Biol. Chem.* **275**, 17201–17210 (2000).
46. Bouvier, M. et al. Removal of phosphorylation sites from the β_2-adrenergic receptor delays onset of agonist-promoted desensitization. *Nature* **333**, 370–373 (1988).
47. Krasel, C., Bünemann, M., Lorenz, K. & Lohse, M. J. β-Arrestin binding to the β2-adrenergic receptor requires both receptor phosphorylation and receptor activation. *J. Biol. Chem.* **280**, 9528–9535 (2005).
48. Merrifield, C. J., Feldman, M. E., Wan, L. & Almers, W. Imaging actin and dynamin recruitment during invagination of single clathrin-coated pits. *Nat. Cell Biol.* **4**, 691–698 (2002).
49. Stauffer, T. P., Ahn, S. & Meyer, T. Receptor-induced transient reduction in plasma membrane PtdIns(4,5)P2 concentration monitored in living cells. *Curr. Biol.* **8**, 343–346 (1998).
50. Schneider, C. A., Rasband, W. S. & Eliceiri, K. W. NIH Image to ImageJ: 25 years of image analysis. *Nat. Methods* **9**, 671–675 (2012).
51. Schindelin, J. et al. Fiji: an open-source platform for biological-image analysis. *Nat. Methods* **9**, 676–682 (2012).
52. Nair, D. et al. Super-resolution imaging reveals that AMPA receptors inside synapses are dynamically organized in nanodomains regulated by PSD95. *J. Neurosci.* **33**, 13204–13224 (2013).
53. Rossier, O. et al. Integrins β1 and β3 exhibit distinct dynamic nanoscale organizations inside focal adhesions. *Nat. Cell Biol.* **14**, 1057–1067 (2012).
54. Humphrey, W., Dalke, A. & Schulten, K. VMD: visual molecular dynamics. *J. Mol. Graph.* **14**, 27–38 (1996).

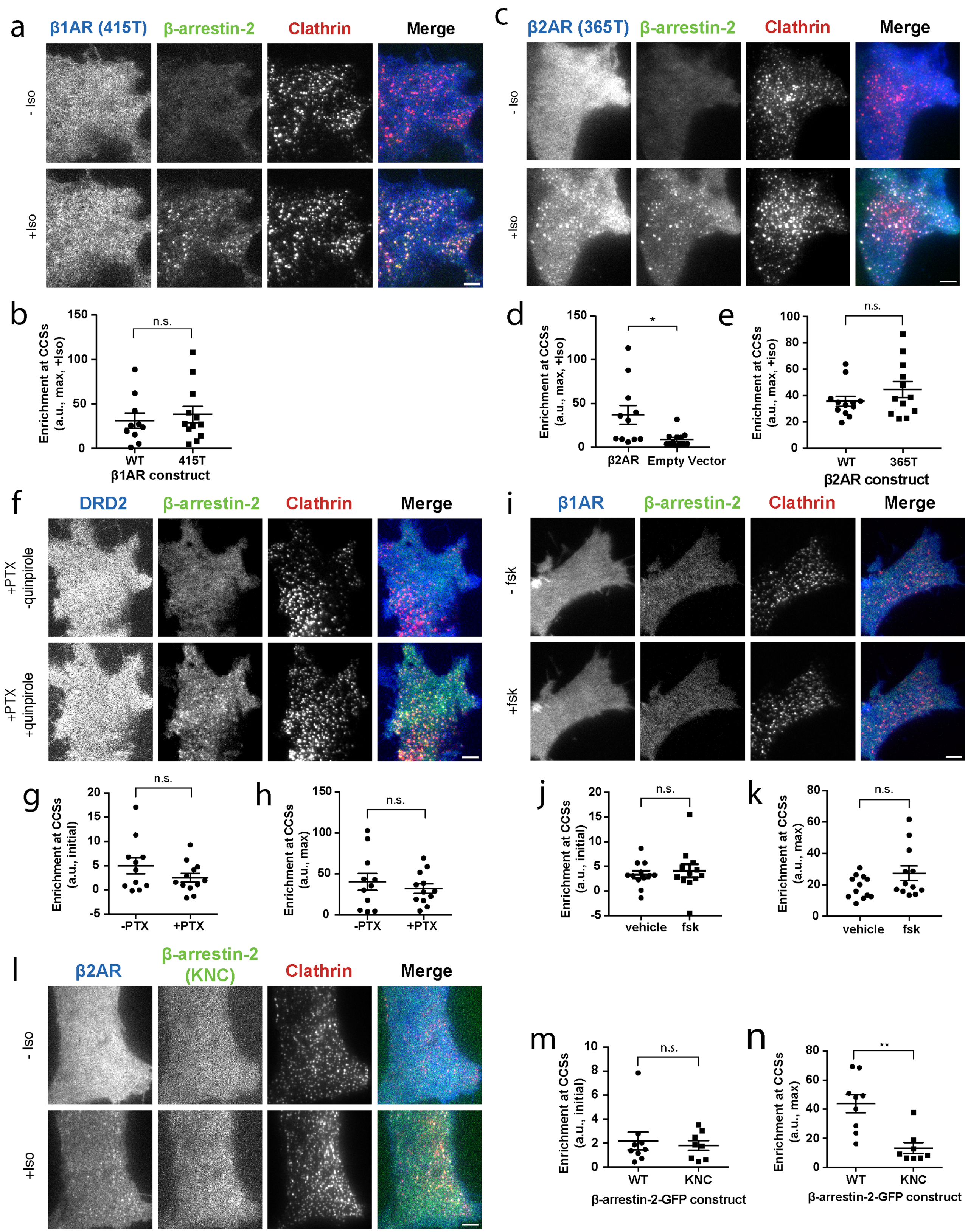


Extended Data Fig. 1 | See next page for caption.

Extended Data Fig. 1 | Verification of GPCR-specificity of the discrete β-arrestin trafficking mechanism, demonstration that this mechanism produces super-stoichiometric β-arrestin accumulation in CCSs and that its activation does not require the GPCR tail. **a**, Mean β-arrestin-2–GFP enrichment at CCSs in cells expressing Flag–β1AR after the following treatments: 10 μM isoproterenol (green, *n* = 14 cells), 15-min pretreatment with 10 μM CGP 20712 A and treatment with 10 μM isoproterenol (red, *n* = 12 cells), or 10 μM CGP 20712 A alone (grey, *n* = 12 cells). Data shown for the 10 μM isoproterenol condition are replotted from Fig. 1b. **b**, Maximum β-arrestin-2–GFP enrichment at CCSs in HEK 293 cells transfected with the indicated receptor or empty vector and treated with 10 μM isoproterenol. **c**, β-arrestin-2–GFP enrichment at CCSs in H9c2 cells without GPCR overexpression and treated with 10 μM isoproterenol or 10 μM dobutamine (*n* = 5 or 4 cells, respectively, from two independent experiments). **d**, β-arrestin-2–GFP enrichment at CCSs in H9c2 cells without GPCR overexpression and treated as indicated (*n* = 12 cells). **e**, Live-cell TIRF microscopy images (representative of *n* = 3 independent experiments) showing Flag–β1AR (blue), β-arrestin-1–mVenus (green) and clathrin light chain–DsRed (red) before and after treatment with 10 μM isoproterenol. **f**, Enrichment into CCSs (*n* = 7 cells from three independent experiments). **g**, Maximum β-arrestin-1–mVenus enrichment at CCSs in HEK 293 cells transfected with Flag–β1AR or empty vector and treated with 10 μM isoproterenol (*n* = 7 and 11 cells, respectively, from three independent experiments; $P = 0.0023$ using an unpaired *t*-test with Welch's correction). **h**, Mean β-arrestin-2–GFP enrichment at CCSs in cells expressing Flag–β2AR after the following treatments: 10 μM isoproterenol (green, *n* = 15 cells), 15-min pretreatment with 10 μM ICI 118,551 and then treatment with 10 μM isoproterenol (red, *n* = 14 cells) or 10 μM ICI 118,551 (grey, *n* = 12). Data shown for the 10 μM isoproterenol condition are replotted from Fig. 1d. **i**, Fluorescence intensity profiles from lines shown in Fig. 1e. **j**, Time-dependent correlation coefficient of line scans across cells derived from immobilization experiments shown in Fig. 1e, f (*n* = 3). **k**, Live-cell TIRF microscopy images (representative of *n* = 3 independent experiments) showing Flag–β2AR–GFP and clathrin-light-chain–DsRed (red) before and after treatment with 10 μM isoproterenol. Fluorescence from the Alexa Fluor 647-conjugated Flag antibody shown in blue and GFP fluorescence shown in green. **l**, Difference in GFP and Alexa Fluor 647 fluorescence enrichment at CCSs in cells co-expressing Flag–β1ARs (red), Flag–β2ARs (blue) and β-arrestin-2–GFP or Flag–β2AR-GFP (black). Cells were labelled with Alexa Fluor 647-conjugated Flag antibody for 10 min before live-cell imaging. Data were derived from the experiments shown in Fig. 1a, b (blue line, *n* = 14 cells from three independent experiments), Fig. 1c, d (red line, *n* = 15 cells from three independent experiments) and Extended Data Fig. 1k (black line *n* = 12 cells from three independent experiments). **m**, Plot of β-arrestin–GPCR stoichiometry calculated from the data displayed in **k**, calibrated according to the double-labelled Flag–β2AR–GFP reference construct, which defines 1:1 stoichiometry (For β1AR and β2AR, *n* = 14 and 15 cells, respectively, from three independent experiments). A correction index was calculated by dividing GFP fluorescence by Alexa Fluor 647 (Flag) fluorescence in CCSs. This correction index was then applied to receptor and β-arrestin-2 enrichment in CCSs to determine β-arrestin-2–GPCR stoichiometry throughout the time course. Images were captured continuously at 0.5 Hz and stoichiometry values over the time course were calculated using a rolling average with 50-frame window size. Scale bar, 5 μm. Scatter plots show overlay of mean and s.e.m. **a**, **d**, **h**, **l** and l show data as mean ± s.e.m. $^{**}P < 0.01$.

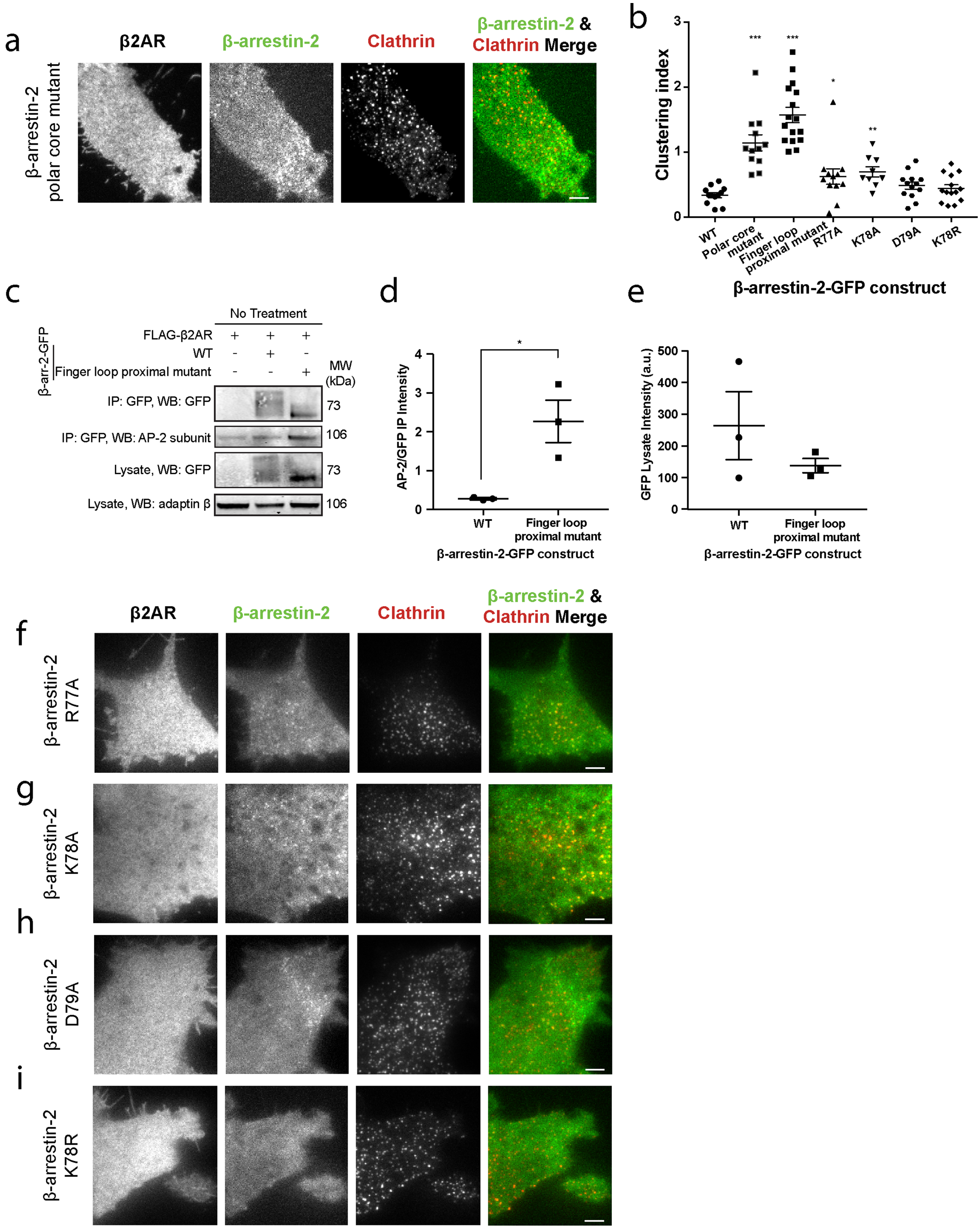


Extended Data Fig. 2 | Additional demonstration that multiple GPCRs can activate the discrete β-arrestin trafficking mechanism. **a**, Live-cell TIRF microscopy images showing Flag–MOR (blue), β-arrestin-2–GFP (green) and clathrin light chain–DsRed (red) before and after treatment with the opioid peptide DAMGO (10 μM). **b**, Mean Flag–MOR and β-arrestin-2–GFP enrichment at CCSs after treatment with 10 μM DAMGO ($n = 12$ cells). **c**, Maximum β-arrestin-2–GFP enrichment at CCSs for HEK 293 cells expressing Flag–MOR or empty vector and treated with 10 μM DAMGO ($n = 12$ cells per condition from three independent experiments; $P < 0.0001$ using a two-tailed unpaired *t*-test with Welch's correction). **d**, Live-cell TIRF microscopy images showing Flag–KOR (blue), β-arrestin-2–GFP (green) and clathrin light chain–DsRed (red) before and after treatment with 10 μM dynorphin. **e**, Enrichment into CCSs after bath application of 10 μM dynorphin ($n = 18$ cells). **f**, Maximum β-arrestin-2–GFP enrichment at CCSs in HEK 293 cells expressing Flag–KOR or empty vector and treated with 10 μM dynorphin ($n = 18$ and 13 cells, respectively, from three independent experiments; $P = 0.0028$ using a two-tailed unpaired *t*-test with Welch's correction). **g**, Live-cell TIRF microscopy images showing Flag–DRD2 (blue), β-arrestin-2–GFP (green) and clathrin light chain–DsRed (red), before and after treatment with 10 μM quinpirole. **h**, Enrichment into CCSs after bath application of 10 μM quinpirole ($n = 12$ cells). **i**, Maximum β-arrestin-2–GFP enrichment at CCSs in cells expressing Flag–DRD2 or untransfected and treated with 10 μM quinpirole ($n = 11$ and 12 cells from three independent experiments; $P = 0.0095$ using a two-tailed unpaired *t*-test with Welch's correction). **a**, **d** and **g** show representative images from three independent experiments. **b**, **e** and **h** show mean ± s.e.m. Scatter plots show overlay of mean and s.e.m. Scale bars, 5 μm. $^{*}P < 0.05$; $^{**}P < 0.01$; $^{***}P < 0.001$.



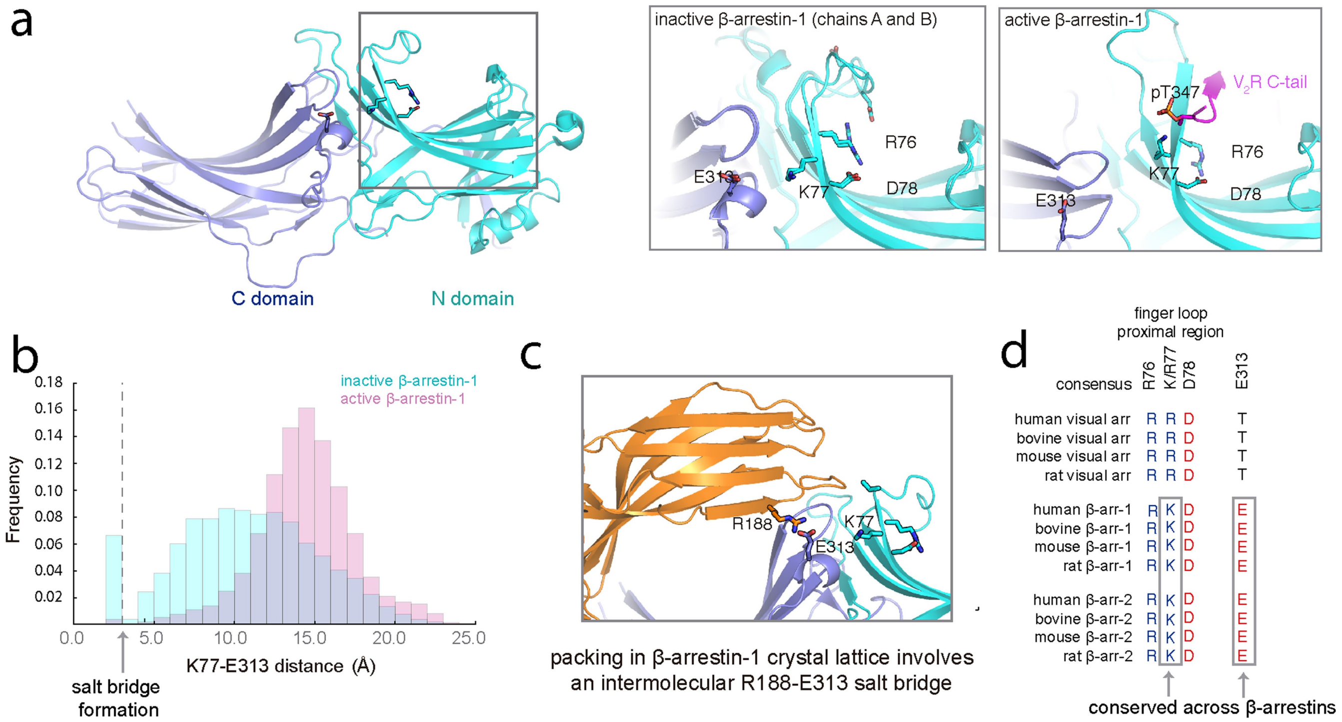
Extended Data Fig. 3 | See next page for caption.

Extended Data Fig. 3 | Direct interaction with the GPCR, but not the GPCR cytoplasmic tail, is required for activation of β-arrestin trafficking. **a**, Live-cell TIRF microscopy images showing Flag–β1AR(415T) (blue), β-arrestin-2–GFP (green) and clathrin light chain–DsRed (red) before and after treatment with 10 μM isoproterenol. **b**, Maximum β-arrestin-2–GFP enrichment at CCSs after treatment with 10 μM isoproterenol for cells co-expressing the indicated Flag–β1AR receptor ($n = 10$ and 12 cells, respectively, from three independent experiments; $P = 0.5825$ calculated using a two-tailed unpaired *t*-test). **c**, Live-cell TIRF microscopy images showing Flag–β2AR(365T) (blue), β-arrestin-2–GFP (green) and clathrin light chain–DsRed (red) before and after treatment with 10 μM isoproterenol. **d**, Maximum β-arrestin-2–GFP enrichment at CCSs in HEK 293 cells treated with 10 μM isoproterenol and transfected with either Flag–β2AR or empty vector ($n = 11$ and 13 cells, respectively, from three independent experiments; $P = 0.0269$ calculated using a two-tailed unpaired *t*-test with Welch's correction). **e**, Maximum β-arrestin-2–GFP enrichment at CCSs for cells co-expressing the indicated Flag–β2AR receptor and treated with 10 μM isoproterenol ($n = 12$ cells from three independent experiments; $P = 0.0606$ calculated using a two-tailed unpaired *t*-test). **f**, Live-cell TIRF microscopy images showing Flag–DRD2 (blue), β-arrestin-2–GFP (green) and clathrin light chain–DsRed (red) before and after treatment with 10 μM quinpirole. **g**, **h**, Initial enrichment in CCSs before treatment with 10 μM quinpirole (**g**) and maximum enrichment after treatment with 10 μM quinpirole (**h**) ($n = 12$ cells from three independent experiments; $P = 0.19$ and 0.4873, respectively, using a two-tailed unpaired *t*-test). **i**, Live-cell TIRF microscopy images showing Flag–β1AR (blue), β-arrestin-2–GFP (green) and clathrin light chain–DsRed (red) before and after treatment with 5 μM forskolin (fsk). **j**, **k**, Initial enrichment in CCSs before treatment with 5 μM forskolin (**j**) and maximum enrichment after treatment with 5 μM forskolin (**k**) ($n = 12$ cells from three independent experiments; $P = 0.6325$ and 0.0971, respectively, using a two-tailed unpaired *t*-test). **l**, Live-cell TIRF microscopy images showing Flag–β2AR (blue), β-arrestin-2–GFP KNC mutant (green) and clathrin light chain–DsRed (red) before and after treatment with 10 μM isoproterenol. **m**, **n**, Initial enrichment in CCSs before treatment with 10 μM isoproterenol (**m**) and maximum enrichment after treatment with 10 μM isoproterenol (**n**) ($n = 9$ (wild-type) or 8 (KNC mutant) cells from three independent experiments; $P = 0.6681$(**m**) and $P = 0.001$ (**n**) using a two-tailed unpaired *t*-test with Welch's correction). **a**, **c**, **f**, **i** and **l** show representative images from three independent experiments. Scatter plots show overlay of mean and s.e.m. Scale bars, 5 μm. *$P < 0.05$, **$P < 0.01$.

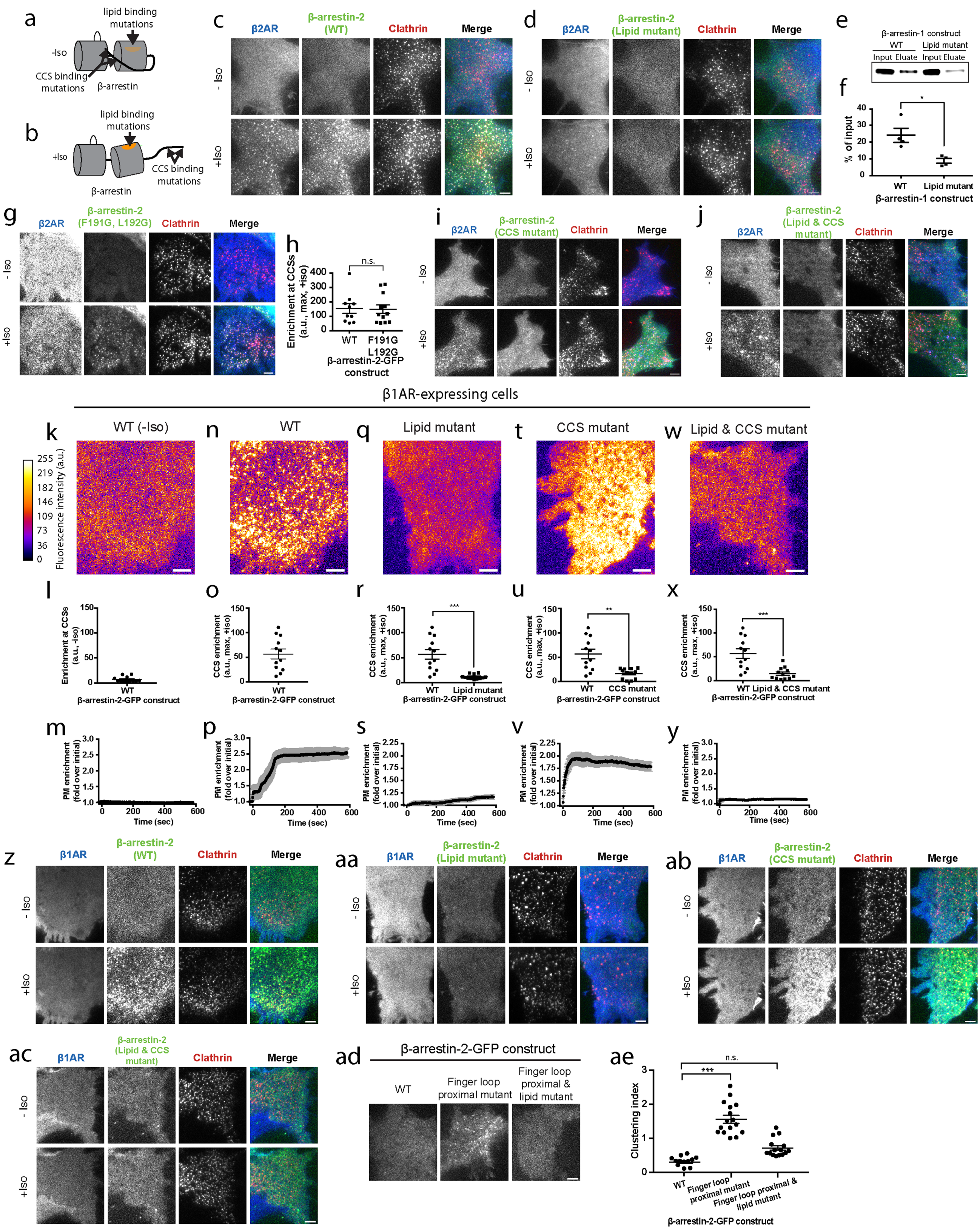


Extended Data Fig. 4 | See next page for caption.

Extended Data Fig. 4 | Additional verification that charge mutations in the finger-loop-proximal region of β-arrestin produce a constitutive activation phenotype. a, Live-cell TIRF microscopy images showing Flag–β2AR, clathrin light chain–DsRed (red) and the polar core mutant of β-arrestin-2–GFP (green) in the absence of agonist treatment. **b**, Clustering index of β-arrestin-2–GFP for the indicated construct in the absence of agonist treatment. Statistical significance was calculated using a two-tailed unpaired *t*-test with Welch's correction (polar core mutant: $n = 12$ cells from three independent experiments, $P < 0.0001$; finger-loop-proximal mutant: $n = 16$ cells from three independent experiments, $P < 0.0001$; R77A: $n = 12$ cells from three independent experiments, $P = 0.0403$; K78A: $n = 12$ cells from three independent experiments, $P = 0.0016$). Wild-type and finger-loop-proximal mutant data is replotted from Fig. 3b. **c**, Association of β-arrestin-2–GFP constructs with the β-adaptin subunit of AP-2 in the absence of agonist treatment. Molecular mass markers (in kDa) are shown on the right side of blots. For gel source data, see Supplementary Fig. 1. The western blots in **c** are representative of three independent experiments, quantified in **d**, and shown as AP-2/GFP intensity in the immunoprecipitation conditions ($n = 3$ independent experiments, $P = 0.0218$ using a two-tailed unpaired *t*-test). **e**, β-arrestin-2–GFP expression in cell lysates from **c**. **f**–**i**, Live-cell TIRF microscopy images showing Flag–β2AR, clathrin light chain–DsRed (red) and β-arrestin-2–GFP with the indicated point mutations (green) in the absence of agonist treatment. Detailed descriptions of β-arrestin mutations are provided in Extended Data Table 1. **a** and **f**–**i** show representative images from three independent experiments. Scatter plots show overlay of mean and s.e.m. Scale bars, 5 μm. *$P < 0.05$, **$P < 0.01$, ***$P < 0.001$.

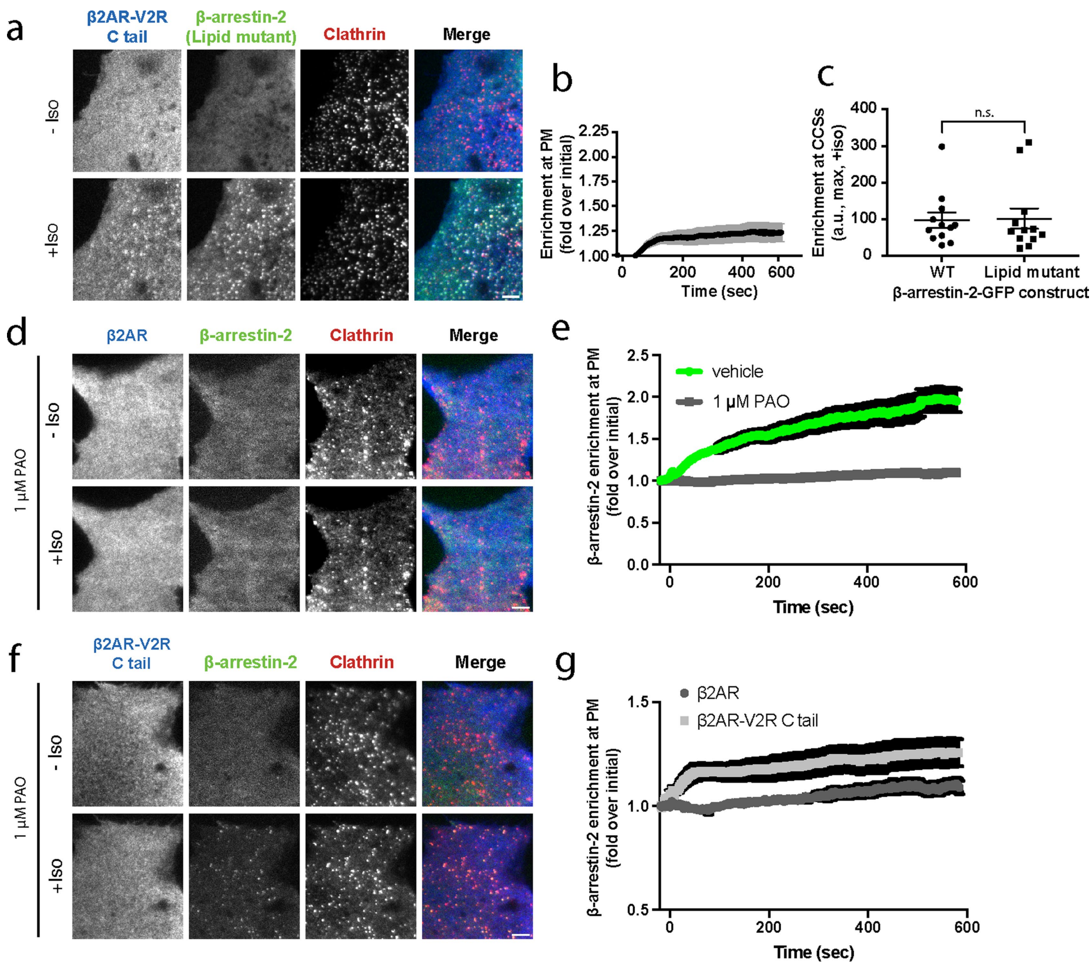


Extended Data Fig. 5 | Molecular dynamics simulations suggest that finger-loop-proximal charged residues stabilize β-arrestin in an inactive state. a, Crystal structures of β-arrestin-1 (left) in an inactive (middle) and active (right) conformation reveal an extensive network of polar residues proximal to the finger loop involving residues R76, K77 and D78. **b**, Histogram of distances between K77 and E313 in simulations of inactive β-arrestin-1 (blue) and active β-arrestin-1 (pink), showing frequency of K77–E313 salt-bridge formation. The K77–E313 distance corresponds to the minimum distance between polar heavy atoms on the two residues' side chains. A separation distance of less than 3.0 Å corresponds to formation of the salt bridge. For the six simulations started from the inactive state, the salt bridge formed 1.1%, 5.7%, 6.3%, 17.6%, 1.0% and 2.0% of the time, respectively (simulation lengths were 4.7 µs, 3.1 µs, 2.9 µs, 5.1 µs, 5.2 µs and 5.7 µs, respectively). For the six simulations started from the active state, the salt bridge formed 0.02%, 0.04%, 0.0%, 1.1%, 0.0% and 0.0% of the time, respectively (simulation lengths were 5.0 µs, 5.0 µs, 4.7 µs, 4.8 µs, 4.8 µs and 5.0 µs, respectively). **c**, Inactive-state crystal structure of β-arrestin-1 in which E313 interacts with R188 on a different β-arrestin-1 molecule in the crystal lattice. **d**, Sequence alignment of arrestins showing conservation of residues R76, K77, D78 and E313. Detailed description of β-arrestin mutations are provided in Extended Data Table 1.

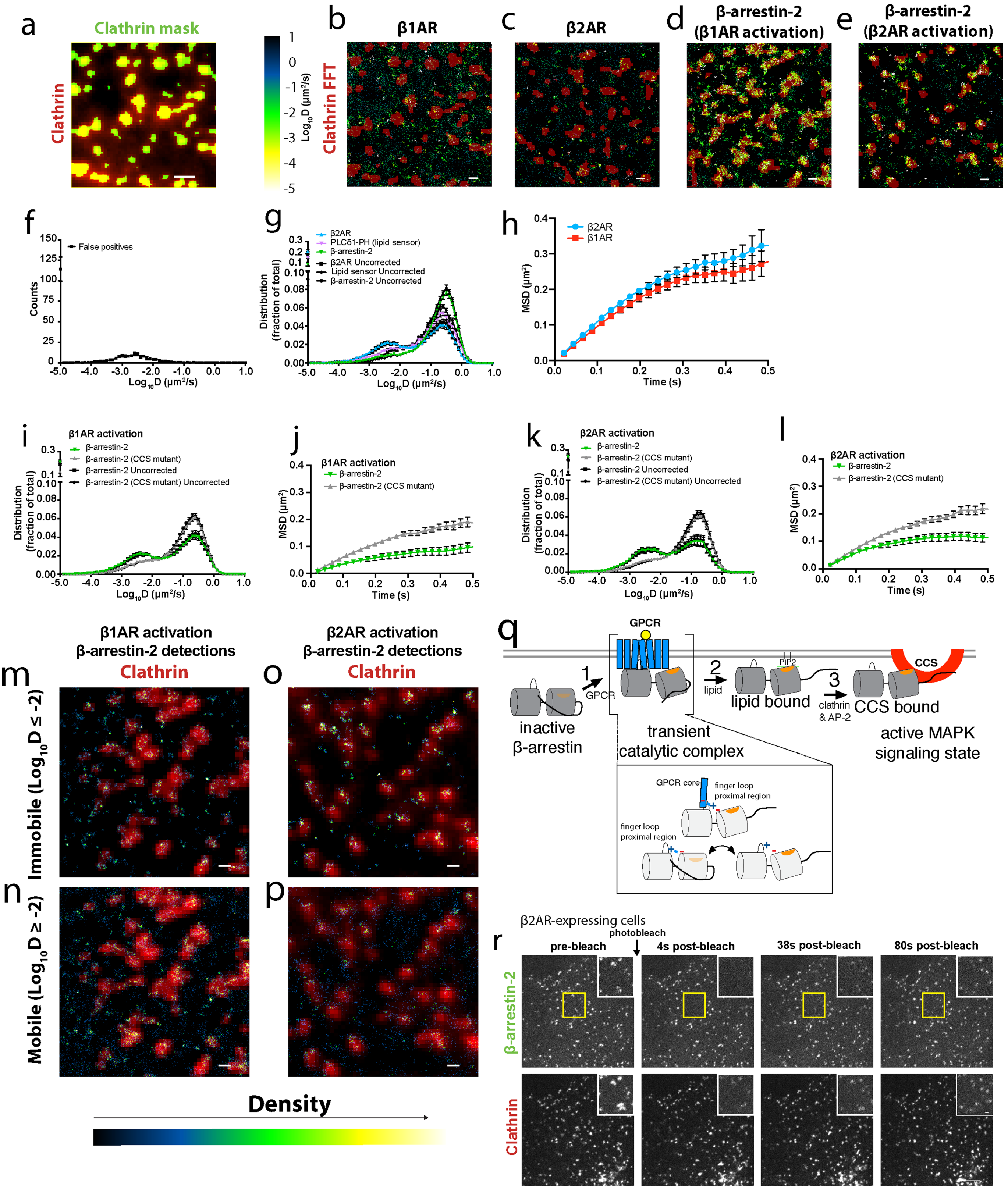


Extended Data Fig. 6 | See next page for caption.

Extended Data Fig. 6 | Verification that the conserved phosphoinositide binding determinant in the β-arrestin C domain is specifically required for the catalytic trafficking mechanism and operates upstream of clathrin- and AP-2-binding interactions. a, **b**, Graphical representation of β-arrestin interaction domains without (**a**) and with (**b**) βAR activation by isoproterenol. **c**, Live-cell TIRF microscopy images showing Flag–β2AR (blue), β-arrestin-2–GFP (green) and clathrin light chain–DsRed (red) before and after treatment with 10 μM isoproterenol. **d**, Live-cell TIRF microscopy images showing Flag–β2AR (blue), β-arrestin-2–GFP lipid mutant (green) and clathrin light chain–DsRed (red) before and after treatment with 10 μM isoproterenol. **e**, Representative western blot (from four independent experiments) of purified wild-type and lipid mutant versions of β-arrestin-1(1–393) immunoprecipitation with PtdIns(4,5)P_2-coated agarose beads and quantified in **f** as per cent of input protein ($n = 4$ independent experiments, $P = 0.0142$ using a two-tailed unpaired *t*-test). For gel source data, see Supplementary Fig. 1. **g**, Live-cell TIRF microscopy images showing Flag–β2AR (blue), β-arrestin-2(F191G/L192G)–GFP lipid anchor mutant (green), and clathrin light chain–DsRed (red) before and after treatment with 10 μM isoproterenol. **h**, Maximum β-arrestin-2–GFP enrichment at CCSs in cells expressing the indicated β-arrestin-2–GFP construct and treated with 10 μM isoproterenol ($n = 12$ cells from three independent experiments; $P = 0.9227$ calculated using a two-tailed unpaired *t*-test). **i**, Live-cell TIRF microscopy images showing Flag–β2AR (blue), β-arrestin-2–GFP CCS mutant (green) and clathrin light chain–DsRed (red) before and after treatment with 10 μM isoproterenol. **j**, Representative images of HEK 293 cells co-expressing Flag–β2AR (blue), β-arrestin-2–GFP lipid and CCS mutant (green) and clathrin light chain–DsRed (red) before and after treament with 10 μM isoproterenol. Representative β-arrestin images false coloured to indicate fluorescence intensity, maximum fluorescence enrichment at CCSs and normalized average plasma membrane (PM) β-arrestin-2–GFP fluorescence (mean ± s.e.m.), respectively, from cells co-expressing Flag–β1ARs ($n = 12$ cells per condition) without isoproterenol treatment (**k**–**m**), and the following β-arrestin-2–GFP constructs with treatment with 10 μM isoproterenol: wild-type (**n**–**p**), lipid mutant (**q**–**s**), CCS mutant (**t**–**v**) and CCS and lipid mutant (**w**–**y**). Wild-type β-arrestin-2–GFP maximum enrichment at CCSs shown in **r**, **u** and **x** is replotted from **o**. **z**–**ac**, Live-cell TIRF microscopy images showing cells before and after treatment with 10 μM isoproterenol and co-expressing Flag–β1AR (blue), clathrin light chain–DsRed (red) and the following GFP-labelled versions of β-arrestin-2 (green): wild-type (**z**), lipid mutant (**aa**), CCS mutant (**ab**), and CCS and lipid mutant (**ac**). **ad**, Live-cell TIRF microscopy images showing Flag–β2AR and the indicated β-arrestin-2–GFP construct in the absence of agonist treatment. **ae**, Clustering index of β-arrestin-2–GFP for the indicated construct in the absence of agonist treatment. Detailed description of β-arrestin mutations are provided in Extended Data Table 1. **c**, **d**, **g**, **i**–**k**, **n**, **q**, **t**, **w** and **z**–**ad** show representative images from three independent experiments. In **r**, **u** and **x**, $n = 12$ cells from three independent experiments; unpaired *t*-test with Welch's correction, $P = 0.0007$, 0.0018 and 0.0012, respectively. In **ae**, $n = 12$ (wild-type) and 16 (finger-loop-proximal mutant) from three independent experiments; unpaired *t*-test with Welch's correction, $P < 0.0001$; $n = 12$ (wild-type) and 15 (finger loop proximal & lipid mutant) from three independent experiments; $P = 0.5464$. Wild-type and finger-loop-proximal mutant data replotted from Fig. 3b. Scatter plots show overlay of mean and s.e.m. Scale bars, 5 μm. $^{**}P < 0.01$, $^{***}P < 0.001$.

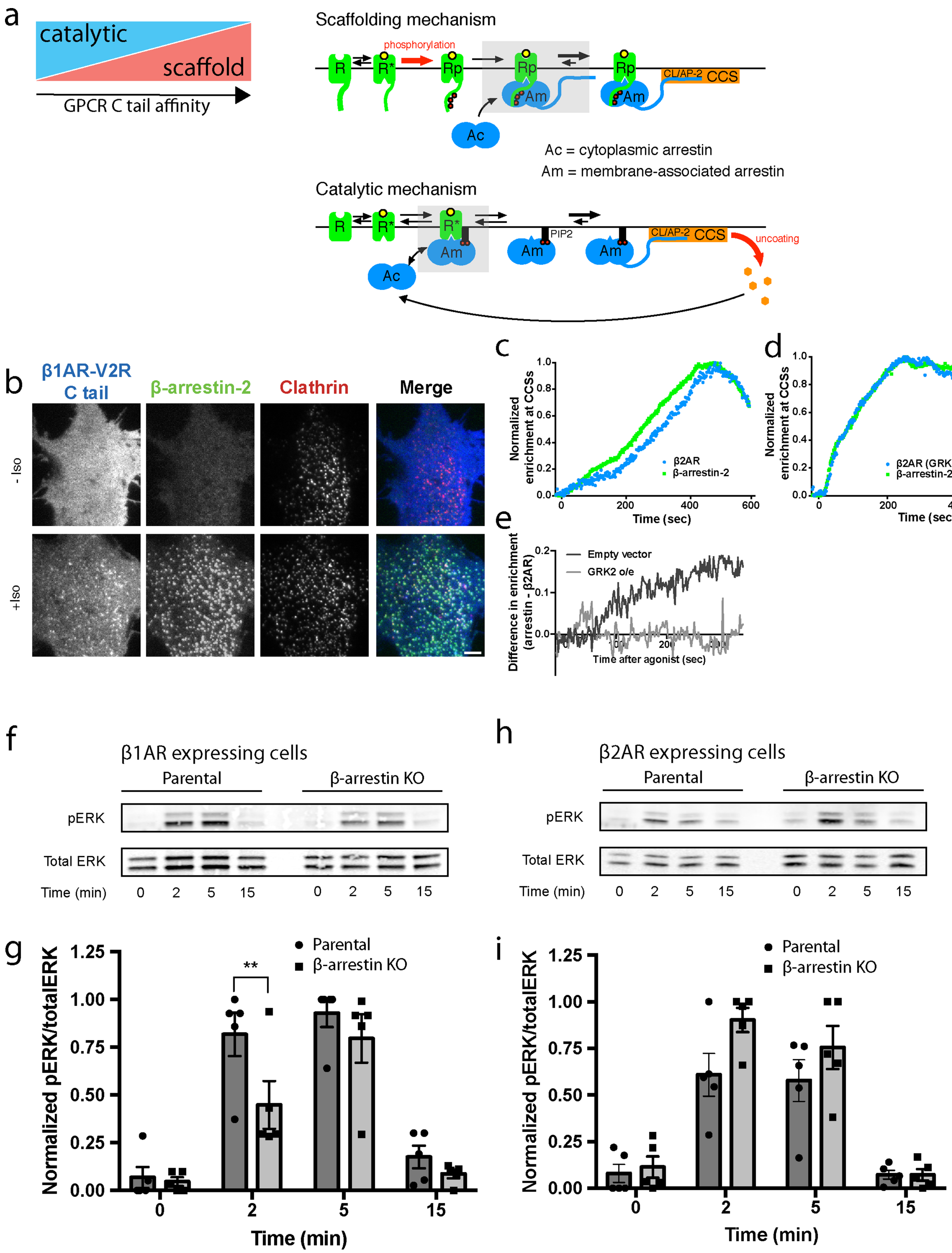


Extended Data Fig. 7 | Phosphoinositide binding is essential for catalytic activation of β-arrestin trafficking but is dispensable for trafficking mediated by the scaffold mechanism. a, Live-cell microscopy images of HEK 293 cells co-expressing Flag–β2AR–V2R C tail (blue), β-arrestin-2–GFP CCS mutant (green) and clathrin light chain–DsRed (red) before and after treatment with 10 μM isoproterenol. **b**, Normalized plasma membrane fluorescence of β-arrestin-2–GFP lipid mutant in cells co-expressing Flag–β2AR–V2R ($n = 12$ cells from three independent experiments) when treated with 10 μM isoproterenol. **c**, Maximum β-arrestin-2–GFP enrichment at CCSs in cells expressing indicated β-arrestin-2–GFP before and after activation of Flag–β2AR–V2R chimaera with 10 μM isoproterenol ($n = 10$ and 12 cells, respectively, from three independent experiments; $P = 0.6433$, two-tailed unpaired *t*-test). **d**, Live-cell microscopy images of COS-1 cells co-expressing Flag–β2AR (blue), β-arrestin-2–GFP (green) and clathrin light chain–DsRed (red) that have been pre-treated for 1 h with 1 μM PAO or vehicle (DMSO) before treatment with 10 μM isoproterenol. **e**, Normalized mean fold increase over initial β-arrestin-2–GFP fluorescence in cells co-expressing Flag–β2AR when pre-treated for 1 h with 1 μM PAO following treatment with 10 μM isoproterenol ($n = 12$ cells from three independent experiments). **f**, Live-cell microscopy images of COS-1 cells co-expressing Flag–β2AR–V2R C tail (blue), β-arrestin-2–GFP (green), and clathrin light chain (CLC)–dsRed (red) that have been pre-treated for 1 h with 1 μM PAO before before treatment with 10 μM isoproterenol. **g**, Normalized average fold increase over initial β-arrestin-2–GFP fluorescence in cells co-expressing Flag–β2AR or Flag–β2AR–V2R when pre-treated for 1 h with 1 μM PAO following treatment with 10 μM isoproterenol ($n = 12$ cells from three independent experiments). **a**, **d** and **f** show representative images from three independent experiments. **b**, **e** and **g** show mean ± s.e.m. Scatter plots show overlay of mean and s.e.m.



Extended Data Fig. 8 | See next page for caption.

Extended Data Fig. 8 | sptPALM controls, MSD plots and cellular model. **a**, Representative image of a clathrin mask (green) generated from a CLC–GFP image (red). **b**–**e**, Representative diffusion maps overlaid with the clathrin mask for HEK 293 cells treated with 10 μM isoproterenol and expressing PAmCherry–β1AR (**b**), PAmCherry–β2AR (**c**), β-arrestin-2–PAmCherry and Flag–β1AR (**d**), and β-arrestin-2–PAmCherry and Flag–β2AR (**e**). **f**, Distribution of diffusion coefficients (*D*) of false-positive detections from HEK 293 cells expressing Flag–β2AR and imaged under standard sptPALM acquisition conditions to determine contribution of false-positive detections in the experimental setup and analysis. **g**, Distribution of diffusion coefficients of PAmCherry–β2AR, PAmCherry–PLCδ1-PH and β-arrestin-2–PAmCherry in live cells imaged at 37 °C after treatment with 10 μM isoproterenol ($n = 13$, 21 and 8 cells, respectively). Black lines show diffusion coefficient profiles that have not been corrected for false-positive detections, showing limited contribution to the profiles. β-arrestin-2–PAmCherry and PAmCherry–PLCδ1-PH were co-expressed individually with Flag–β2AR. **h**, Average MSD plots derived from sptPALM analysis of PAmCherry–β1AR and PAmCherry–β2AR trajectories in HEK 293 cells treated with 10 μM isoproterenol ($n = 8$ and 13 cells, respectively). **i**, Distribution of diffusion coefficients of wild-type and CCS mutant β-arrestin-2–PAmCherry when co-expressed with Flag–β1AR in live HEK 293 cells imaged at 37 °C after treatment with 10 μM isoproterenol ($n = 13$ and 17 cells, respectively). Black lines show diffusion coefficient profiles that have not been corrected for false-positive detections, showing limited contribution to the profiles. **j**, Average MSD plots derived from sptPALM analysis of wild-type and CCS mutant β-arrestin-2–PAmCherry trajectories in cells co-expressing Flag–β1AR and treated with 10 μM isoproterenol ($n = 13$ and 17 cells, respectively). **k**, Distribution of diffusion coefficients of wild-type and CCS mutant β-arrestin-2–PAmCherry when co-expressed with Flag–β2AR in live cells imaged at 37 °C after treatment with 10 μM isoproterenol ($n = 21$ and 10 cells, respectively). Black lines show diffusion coefficient profiles that have not been corrected for false-positive detections, showing limited contribution to the profiles. β-arrestin-2–PAmCherry diffusion coefficient profiles when activated by the β2AR are replotted from **g**. **l**, Average MSD plots derived from sptPALM analysis of wild-type and CCS mutant β-arrestin-2–PAmCherry trajectories in HEK 293 cells co-expressing Flag–β2AR and treated with 10 μM isoproterenol ($n = 21$ and 10 cells, respectively). **m**, **n**, Immobile (**m**) and mobile(**n**) β-arrestin-2–PAmCherry trajectory detections overlaid with a clathrin marker (red) in live cells co-expressing Flag–β1AR after treatment with 10 μM isoproterenol. **o**, **p**, Immobile (**o**) and mobile (**p**) β-arrestin-2–PAmCherry trajectory detections overlaid with a clathrin marker (red) in live cells co-expressing Flag–β2AR after treatment with 10 μM isoproterenol. Trajectory detections are false coloured based on the density of detections at each pixel. Error bars represent s.e.m; in some cases, error bars are smaller than the height of the symbol and are therefore not shown. Scale bars, 500 nm for sptPALM images. **q**, Proposed cellular pathway for catalytic activation of β-arrestin. **r**, Representative microscopy images of COS-1 cells co-expressing Flag–β2AR, β-arrestin-2–GFP (green) and clathrin light chain–DsRed (red) that were treated with 10 μM isoproterenol for 3 min. Then, β-arrestin-2–GFP was photobleached in the indicated yellow region (inset; insets are also shown in Fig. 5h). **a**–**e, m**–**p** and **r** show representative examples from at least three independent experiments. **f**–**l** show data as mean ± s.e.m.; in some cases, error bars are smaller than the height of the symbol and are, therefore, not shown. Scale bars; 500 nm for sptPALM images, 5 μm for larger FRAP images, 0.5 μm for inset FRAP images.



Extended Data Fig. 9 | See next page for caption.

Extended Data Fig. 9 | Differences in the bioenergetics of catalytic versus scaffold mechanisms of regulated β-arrestin trafficking and β-arrestin-dependent activation of ERK1/2 promoted by catalytic activation. a, Schematic depicting the proposed co-existence of catalytic and scaffolding mechanisms of β-arrestin trafficking tuned according to tail binding affinity, emphasizing the difference in tail versus core interactions (shaded boxes). The tail interaction, which requires GPCR phosphorylation (Rp), drives scaffold formation by stabilizing GPCR–β-arrestin complex formation. The core interaction mediates catalysis by providing a kinetically favourable path for β-arrestin to remain at the plasma membrane irrespective of whether it is associated with a GPCR. The core interaction requires phosphoinositide binding to the β-arrestin C domain, which explains why the phosphoinositide requirement is specific to the catalytic mechanism and can be overcome by formation of a sufficiently stable scaffold complex requiring the phosphorylated GPCR tail. Primary energy inputs maintaining each proposed trafficking cycle are indicated by red arrows. The present results identify a specific requirement of the catalytic mechanism for phosphoinositide binding to the C domain but they do not exclude binding in the scaffold complex (which we think is likely). We also cannot presently rule out the possible existence of additional interaction(s) in the catalytic mechanism, such as phosphoinositide binding to the β-arrestin N domain, which has the potential to displace the β-arrestin C terminus[22]. **b**, Representative TIRF microscopy images of live cells (from three independent experiments) before and after treatment of cells expressing chimaeric Flag-tagged β1AR-V2Rs with 10 μM isoproterenol. Profiles of Flag-β2AR and β-arrestin-2–GFP average enrichment into CCSs in COS-1 cells expressing either an empty vector construct (**c**) or GRK2 (**d**) and treated with 10 μM isoproterenol ($n = 15$ or 12 cells, respectively, from three independent experiments). **e**, Difference in enrichment between β-arrestin-2–GFP and β2AR in cells shown in **c** and **d**, showing the effect of GRK2 overexpression. **f**, Representative western blot showing phosphorylated ERK1/2 and total ERK1/2 signal in extracts prepared from parental or β-arrestin-CRISPR-knockout HEK 293 cells expressing Flag–β1AR and exposed to 10 μM isoproterenol for the indicated time period. **g**, Quantification of ERK1/2 activation from the western blots in **f** ($n = 5$ independent experiments, $P = 0.004$ using a one-way ANOVA). **h**, Representative western blot showing phosphorylated ERK1/2 and total ERK1/2 signal in extracts prepared from parental or β-arrestin-CRISPR-knockout HEK 293 cells expressing Flag–β2AR and exposed to 10 μM isoproterenol for the indicated time period. **i**, Quantification of ERK1/2 activation from the western blots in **h** ($n = 5$ independent experiments). **f** and **h** show representative western blots from five independent experiments. Data are mean ± s.e.m. For gel source data, see Supplementary Fig. 1. Error bars represent s.e.m. $^{**}P < 0.01$.

Extended Data Table 1 | Summary and description of β-arrestin mutations

Mutation name	**Mutations**	**Description**	**Phenotype**
CCS mutant	L373A, I374A, F376A, E375K, E377K, R393A, R395A	Clathrin & AP-2 binding deficient from mutation in C-terminus	Recruits to plasma membrane but not to CCSs
Lipid mutant	K233Q, R237Q, K251Q (K232Q, R236Q, K250Q in β-arrestin-1)	Lipid binding deficient from mutation in C-lobe. Residues not well conserved in visual arrestin.	Not recruited to plasma membrane nor to CCSs
Lipid & CCS mutant	K233Q, R237Q, K251Q, L373A, I374A, F376A, E375K, E377K, R393A, R395A	Lipid, clathrin, & AP-2 binding deficient from mutations in C-lobe and C-terminus	Not recruited to plasma membrane nor to CCSs
Finger loop proximal mutant	R77A, K78A, D79A (R76A, K77A, D78A in β-arrestin-1)	Triple alanine mutations in the region proximal to the finger loop	Constitutively activating – association with CCSs without GPCR activation
K78E	K78E (K77E in β-arrestin-1)	Charge swap mutation of single residue from cluster of finger loop proximal mutations	Constitutively activating – association with CCSs without GPCR activation
E314K	E314K (E313K in β-arrestin-1)	Charge swap mutation of residue in C lobe	Constitutively activating – association with CCSs without GPCR activation
K78E E314K	K78E E314K (K77E E313K in β-arrestin-1)	Double charge swap mutations restoring putative salt bridge	Abrogates constitutively activating phenotype produced by single charge swap mutants
KNC mutant	K11A, K12A, L49A, D51A, R52A, L69A, Y239A, D241A, C252A, P253A, D260A and Q262A	12 alanine mutations disrupting GPCR binding	Not recruited to plasma membrane nor to CCSs
F191G, L192G	F191G, L192G Lipid anchor mutant	Charge mutation of lipid anchor region that is conserved between visual and β-arrestins	Enriched at CCSs comparable to WT
Polar core mutant	D386A, D387A, D388A	Triple alanine mutation in β-arrestin polar core region	Constitutively activating – association with CCSs without GPCR activation

Mutations are in β-arrestin-2 unless otherwise indicated.

https://doi.org/10.1038/s41586-018-0105-3

Reconstitution reveals motor activation for intraflagellar transport

Mohamed A. A. Mohamed[1], Willi L. Stepp[1] & Zeynep Ökten[1,2]*

The human body represents a notable example of ciliary diversification. Extending from the surface of most cells, cilia accomplish a diverse set of tasks. Predictably, mutations in ciliary genes cause a wide range of human diseases such as male infertility and blindness. In *Caenorhabditis elegans* sensory cilia, this functional diversity appears to be traceable to the differential regulation of the kinesin-2-powered intraflagellar-transport (IFT) machinery. Here we reconstituted the first, to our knowledge, functional multi-component IFT complex that is deployed in the sensory cilia of *C. elegans*. Our bottom-up approach revealed the molecular basis of specific motor recruitment to the IFT trains. We identified the key component that incorporates homodimeric kinesin-2 into its physiologically relevant context, which in turn allosterically activates the motor for efficient transport. These results will enable the molecular delineation of IFT regulation, which has eluded understanding since its discovery more than two decades ago.

Having been neglected for decades, the biology of cilia underwent a renaissance after the realization that ciliogenesis is central to development and disease[1–5]. Almost all cilia are built by the highly conserved IFT machinery with a few exceptions in which cilia are built in the cytoplasm in an IFT-independent manner[5–8]. Numerous in vivo studies have established the universal characteristics of IFT in diverse model systems. The IFT trains display continuous movement towards either the ciliary tip (kinesin-2-dependent) or the ciliary base (dynein-2-dependent) without apparent reversals in between[5,7,9–11]. When a kinesin-2-powered IFT train arrives at the ciliary tip, it undergoes a poorly understood remodelling process that deactivates the kinesin-2 motor and restructures the train for dynein-2-dependent transport back to the base[5,7,9–11]. It has long been known that *C. elegans* uses two different kinesin motors, a heterotrimeric kinesin-2 called kinesin-II and a homodimeric kinesin-2 called OSM-3, to build its functionally distinct cilia[12]. Loss of the homodimeric OSM-3 function invariably leads to the loss of the distal portion of the canonical cilium and to deficient osmotic avoidance in the so-called rod-shaped or 'canonical' cilia, whereas the loss of both the heterotrimeric and homodimeric kinesin-2 motors leads to the loss of entire axonemes[12]. In the wing-shaped AWC cilia, on the other hand, loss of the heterotrimeric kinesin-2 motor leads to deficiencies in chemotaxis, suggesting that the kinesin-dependent trafficking of cilia-specific components defines the functional identity of the respective cilia[13,14]. In contrast to the heterotrimeric kinesin-2, the precise role of the homodimeric kinesin-2 in other organisms is much less understood[7]. For instance, the homodimeric kinesin-2 moves along the mammalian cilia, but it has not yet been shown to function as an IFT motor[15,16].

Despite its essential role in ciliogenesis and its emerging role in ciliary diversity[13,14,17–19], the mechanisms of IFT train assembly, motor recruitment, and the timely activation and deactivation of the oppositely directed kinesin-2 and dynein-2 motors are poorly understood. To move towards a comprehensive molecular understanding of IFT, here we used a bottom-up approach to dissect the OSM-3-dependent IFT in *C. elegans* sensory cilia. The OSM-3 function was previously linked to the so-called IFT-B complex[20]. Loss of function in many IFT-B subunits has been proposed to interfere with OSM-3 function in vivo, and many of these subunits have been shown to be part of the IFT-B core complex[20–24] (Extended Data Fig. 1). Previous work with the green algae *Chlamydomonas reinhardtii* was instrumental to the detailed characterization of the IFT components. The entire IFT-B complex consisting of the peripheral and core complexes has been reconstituted from 15 recombinantly expressed subunits[25]. Specifically, nine core components (IFT-74, IFT-81, IFT-27, IFT-25, IFT-22, IFT-52, IFT-46, IFT-88 and IFT-70) form a stable complex; several of these core components can also combine autonomously to form sub-complexes in vitro[26–29]. Guided by these previous findings, we concentrated on the IFT-B core complex and started our bottom-up approach by first assembling the presumptive sub-complexes using recombinantly expressed wild-type, full-length subunits from *C. elegans*. Next, we systematically probed the interaction of the sub-complexes, as well as their individual subunits, with the OSM-3 motor using multiple techniques. To visualize the motor activity in functional-transport assays, we used a previously described, constitutively active OSM-3(G444E) motor[30] that we C-terminally tagged for fluorophore labelling using the HaloTag system (hereafter the tagged motor is referred to as OSM-3(G444E)–Halo).

Building the first IFT–kinesin-2 complex

To monitor complex formation, we made use of size-exclusion chromatography that was coupled to multiple-angle light scattering (SEC–MALS) analysis to gauge the molecular mass of the presumptive complexes. We first assessed whether the tripartite IFT-74–IFT-81–IFT-22(IFTA-2) (symbols in parentheses refer to the *C. elegans* protein nomenclature) core complex (hereafter TCC; blue subunits in Extended Data Fig. 1), and the quadripartite IFT-52(OSM-6)–IFT-46(DYF-6)–IFT-88(OSM-5)–IFT-70(DYF-1) core complex (hereafter QCC; magenta subunits in Extended Data Fig. 1) are formed stably or whether additional factors were necessary (that is, IFT-27 and IFT-25 subunits that are present in the IFT-B core complex of *C. reinhardtii* but not in *C. elegans*[29]). Although both complexes were formed, only the QCC, and not the TCC, incorporated the motor protein (Fig. 1 and Extended Data Fig. 2a–c). Notably, removal of the IFT-70(DYF-1) subunit from the QCC disrupted the complex formation between the motor and its IFT-B complex, suggesting an IFT-70(DYF-1)-specific recruitment of the motor (Fig. 1, middle versus right). Consistent with this notion, the motor was found under the elution peak with QCC

---

[1]Physik Department E22, Technische Universität München, Garching, Germany. [2]Munich Center for Integrated Protein Science, Munich, Germany. *e-mail: zoekten@ph.tum.de

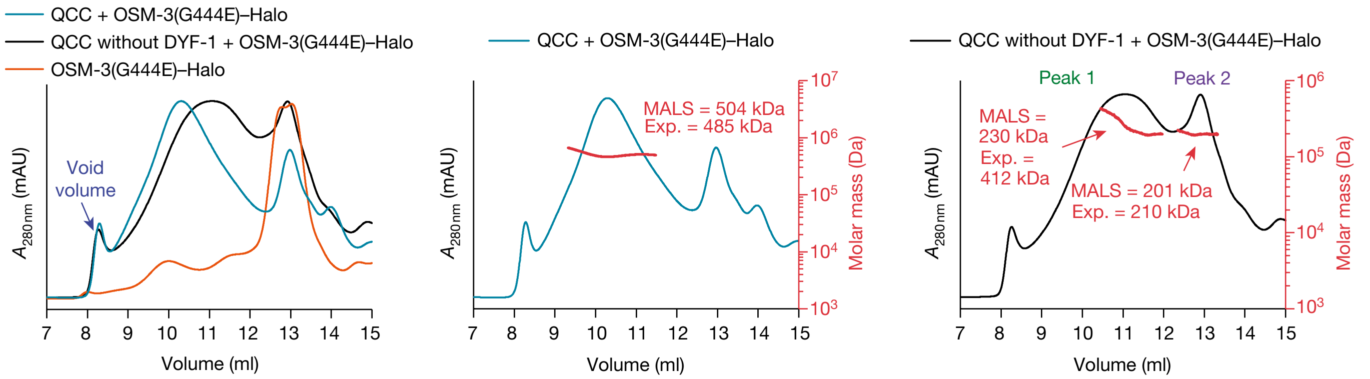


Fig. 1 | IFT-70(DYF-1) is key to the incorporation of the OSM-3 motor into the QCC. Overlay of the elution profiles of QCC lacking IFT-70(DYF-1) (labelled as QCC without DYF-1) + OSM-3(G444E)–Halo and QCC + OSM-3(G444E)–Halo along with the OSM-3(G444E)–Halo motor show the IFT-70(DYF-1)-dependent shift of the elution peaks upon incorporation of the motor into the QCC (left). The MALS analyses of the complexes are also shown (middle and right). The molar mass determined from the MALS fit and the expected mass (Exp.) of the complexes are shown in red (right axes and insets). Results are representative of three independent experiments.

and was absent under QCC peak when the IFT-70(DYF-1) subunit was removed (Extended Data Fig. 2a, middle versus right).

To provide independent biochemical evidence for the IFT-70(DYF-1)-mediated interaction of the motor with the QCC, we turned to in-solution protein-binding assays. Using microscale thermophoresis (MST), we observed robust, sub-micromolar binding of the motor protein to the QCC, whereas removal of the IFT-70(DYF-1) subunit abolished this interaction (Fig. 2). Taken together, our solution-based SEC–MALS and MST analyses demonstrate the specific, IFT-70(DYF-1)-mediated recruitment of the OSM-3 motor to the IFT-B complex.

To further corroborate the key role of the IFT-70(DYF-1) subunit in OSM-3 recruitment, we next turned to fluorescence-based microscopy assays. We functionalized each of the IFT-B subunits with different fluorescent tags and then performed photobleaching assays that detected non-aggregated single subunits (Extended Data Fig. 3). Next, we determined the pairwise colocalization efficiencies by labelling the respective subunits with two different fluorophores. This exclusion process demonstrated that the functionalized subunits retained their capability to assemble into a stable complex under assay conditions (Extended Data Fig. 4). To find out which of these sub-complexes display efficient colocalization with the motor, we differentially labelled OSM-3(G444E)–Halo and each of the sub-complexes. Consistent with our previous assays that demonstrated the IFT-70(DYF-1)-dependent interaction between the motor and its IFT-B complex (Figs. 1, 2), OSM-3(G444E)–Halo failed to colocalize efficiently with TCCs and QCCs that lacked IFT-70(DYF-1), but displayed significant colocalization with the QCC that contained the IFT-70(DYF-1) subunit (Fig. 3a). Finally, the efficient colocalization of the IFT-70(DYF-1) subunit with OSM-3(G444E)–Halo, which did not occur with IFT-52(OSM-6), IFT-88(OSM-5) or IFT-46(DYF-6) subunits, provided direct evidence that IFT-70(DYF-1) is the key subunit that mediates the interaction between the motor and the IFT-B core complex (Fig. 3b). By contrast, the heterotrimeric KLP11–KLP20–KAP motor, which cooperates with OSM-3 to build the rod-shaped sensory cilium in *C. elegans*, failed to colocalize with the IFT-B complex (Extended Data Fig. 5). Previous in vivo experiments suggested that the KLP11–KLP20–KAP motor associates with the IFT-A complex[20].

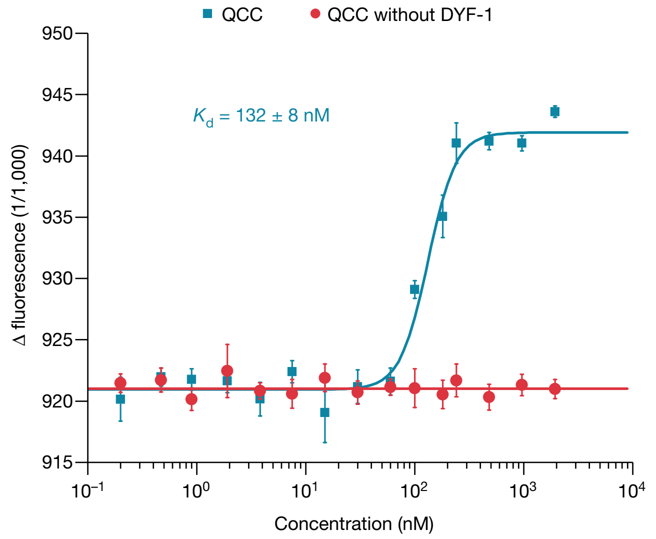


Fig. 2 | IFT-70(DYF-1)-mediated binding of the OSM-3(G444E)–Halo motor to the QCC as measured by MST. The fluorophore-labelled motor was titrated with both the QCC and QCC lacking IFT-70(DYF-1). Data are mean ± s.d. from three independent experiments. Source Data are provided with the online version the paper.

Functional transport assays arguably provide the most direct evidence for the specific recruitment and activation of the OSM-3 motor by the IFT-70(DYF-1) subunit. To assess whether our reconstituted OSM-3–IFT-B complex is capable of directional movement in vitro, we tracked the differentially fluorophore-labelled OSM-3(G444E)–Halo motor and its IFT-B complex using a total internal reflection fluorescence (TIRF) microscope. Most of the fluorophore signals from the motor and its complex were colocalized and moved directionally on surface-attached microtubules (Fig. 4a and Supplementary Video 1). Consistent with our previous assays, removal of IFT-70(DYF-1) again terminated the interaction between the motor with QCC, and the OSM-3(G444E)–Halo motor moved alone along microtubules (Supplementary Video 1). Notably, incorporation of the motor into the QCC further enhanced the velocity of OSM-3(G444E)–Halo, whereas the velocities of the motor alone or in the presence of the IFT-70(DYF-1) subunit were indistinguishable (Fig. 4a). The processivity of the respective motors, by contrast, was independent of the IFT-70(DYF-1) subunit or the QCC (Extended Data Fig. 6a).

The point mutation in the stalk of the OSM-3(G444E)–Halo construct hampers the auto-inhibitory folding, which in turn results in an activated motor in vitro[30]. Therefore, we next asked whether the IFT-70(DYF-1) subunit can also activate the auto-inhibited OSM-3 motor in our reconstitution assays. We designed a motor with a wild-type stalk that lacked the G444E mutation. We introduced an N-terminal Flag and a SNAP-tag for affinity purification and fluorophore labelling of the motor (hereafter OSM-3–SNAP), respectively. In this construct, the entire C terminus that follows the catalytic heads was not modified. As demonstrated with OSM-3(G444E)–Halo (Fig. 3), the OSM-3–SNAP motor with an unmodified C terminus displayed efficient IFT-70(DYF-1)-dependent colocalization with the QCC, but

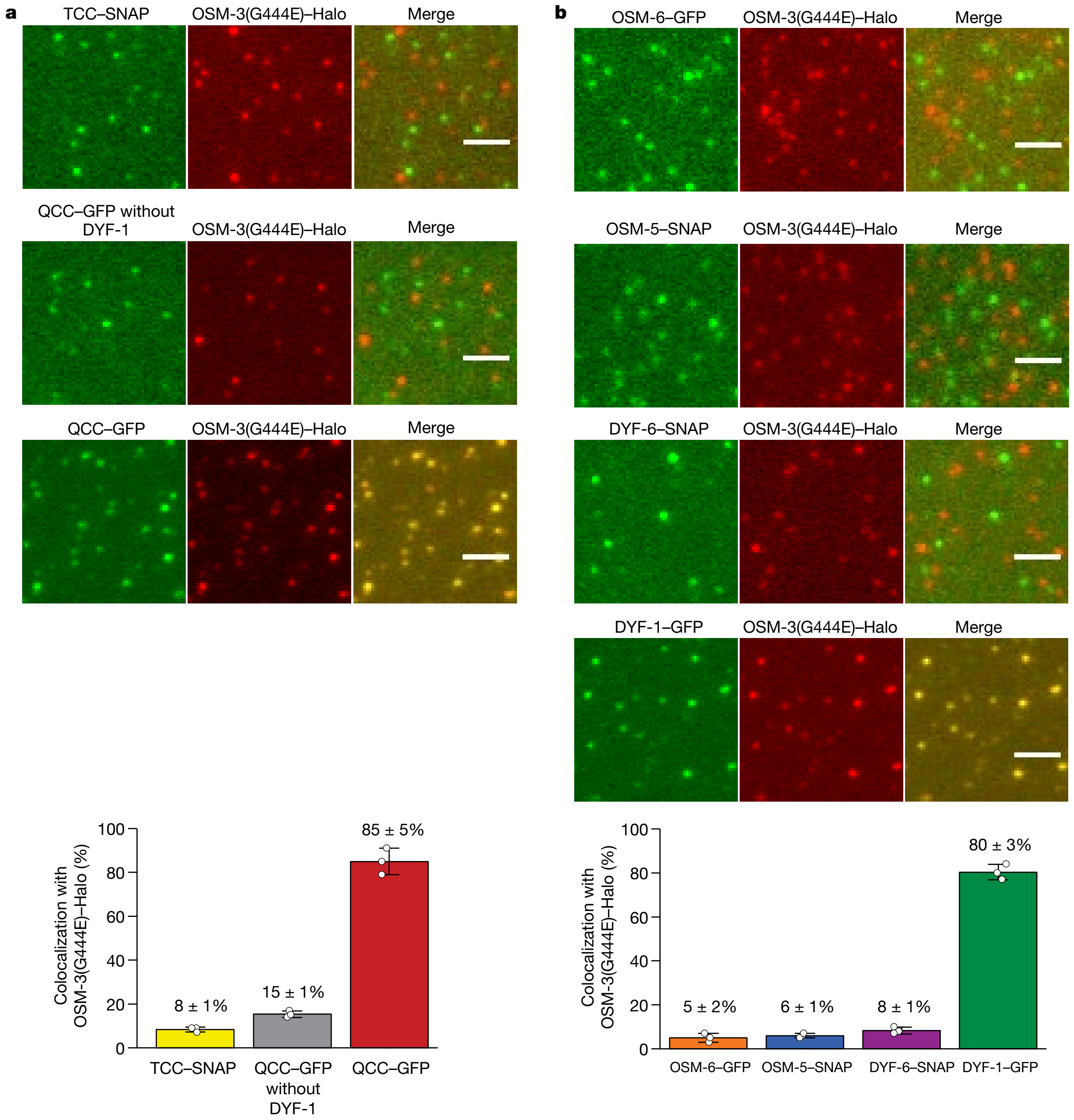


Fig. 3 | Colocalization efficiency of the OSM-3(G444E)–Halo motor with its IFT-B complex is solely dependent on the IFT-70(DYF-1) subunit. a, Neither TCC nor QCC that lacked IFT70(DYF-1) displayed efficient colocalization with the OSM-3(G444E)–Halo motor. In the presence of the IFT-70(DYF-1) subunit, however, the QCC efficiently colocalized with OSM-3(G444E)–Halo (85 ± 5%). **b**, Consistently, OSM-3(G444E)–Halo failed to interact with individual IFT-52(OSM-6), IFT-88(OSM-5) and IFT-46(DYF-6) subunits but showed robust colocalization (80 ± 3%) with the IFT-70(DYF-1) subunit. IFT-81 and IFT-52(OSM-6) were fluorescently labelled with C-terminal SNAP and green fluorescent protein (GFP) tags, respectively. Data are mean ± s.d. from three independent experiments. Source Data are available with the online version of the paper. Scale bars, 3 μm.

not with the TCC (Extended Data Fig. 7a). OSM-3–SNAP interacted with the IFT-70(DYF-1) subunit but not with the IFT-52(OSM-6), IFT-88(OSM-5) or IFT-46(DYF-6) subunits (Extended Data Fig. 7b). However, the auto-regulation in kinesin motors is achieved by folding of the distal tail domain at the C terminus onto the N-terminal head domains[30]. To ensure that the N-terminal SNAP-tag does not impede the auto-inhibitory folding of the motor, we also designed a construct that was only Flag-tagged at its C terminus for affinity purification (hereafter OSM-3–Flag). Owing to the lack of fluorescence information, this motor cannot be followed directly in the functional transport assays; however, it represents the wild-type OSM-3 as closely as possible. Both motors containing the wild-type stalk were recruited and activated in an IFT-70(DYF-1)-specific manner (Fig. 4b, c and Supplementary Videos 2, 3). The IFT-70(DYF-1)-dependent activation was much more obvious with the OSM-3–SNAP motor that lacked the G444E mutation. As expected from an auto-inhibited kinesin, and in stark contrast to the OSM-3(G444E)–Halo motor, OSM-3–SNAP (which has a wild-type stalk) was barely capable of directional movement but rather displayed two-dimensional diffusion (compare the top left of Supplementary Video 1 to the top left of Supplementary Video 2). The presence of IFT-70(DYF-1) or the QCC initiated robust directional movement of the auto-inhibited OSM-3–SNAP motor (Supplementary Video 2). Finally—as with the OSM-3(G444E)–Halo motor (Extended Data Fig. 6a)—once activated by IFT-70(DYF-1), the processivity of OSM-3–SNAP and OSM-3–Flag motors was not modified by the presence of IFT-70(DYF-1) alone or the QCC (Extended Data Fig. 6b, c).

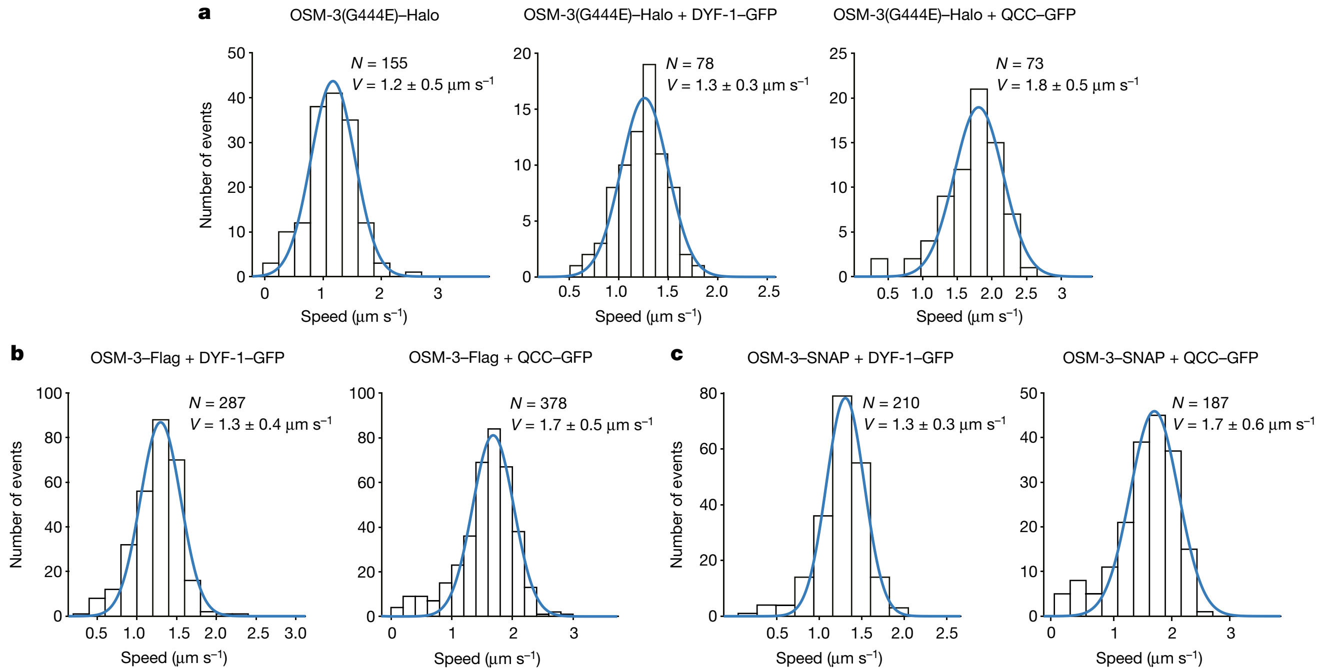


Fig. 4 | IFT-70(DYF-1)-dependent incorporation into the QCC fully activates OSM-3 in vitro. a, In the presence (middle) and absence (left) of IFT-70(DYF-1), the velocity (*V*) of OSM-3(G444E)–Halo is indistinguishable but incorporation into the QCC (right) significantly increases the velocity of the motor. *P* values from a two-tailed *t*-test assuming unequal variances: OSM-3(G444E)–Halo versus OSM-3(G444E)–Halo + DYF-1–GFP, 0.013; OSM-3(G444E)–Halo versus OSM-3(G444E)–Halo + QCC–GFP, 2.5×10^{-18}; OSM-3(G444E)–Halo + DYF-1–GFP versus OSM-3(G444E)–Halo + QCC–GFP, 3.4×10^{-15}. **b**, **c**, Consistently, QCC fully activates the OSM-3–Flag (**b**, $P = 9.6 \times 10^{-25}$) and OSM-3–SNAP (**c**, $P = 1.8 \times 10^{-12}$) complexes but not the IFT-70(DYF-1) subunit alone. *N*, number of events obtained from three different flow chambers in three independent experiments (velocities fit to a Gaussian distribution, mean ± s.d.).

Taken together, our functional transport assays consistently showed that the presence of IFT-70(DYF-1) is essential for the recruitment of the OSM-3 motor to the IFT-B complex. Our results further revealed that the presence of the IFT-70(DYF-1) subunit alone is not sufficient to fully activate the motor. Instead, it is the IFT-70(DYF-1)-mediated incorporation of OSM-3 into the IFT-B complex that allosterically activates the motor. Notably, the transport rates extracted from our in vitro reconstitution assays were consistent with rates extracted from in vivo studies[12,20,31]. The physiological relevance of our reconstitution is underscored by previous in vivo findings in the *C. elegans* sensory cilium. In the absence of *dyf-1* function, OSM-3 not only appeared to detach from the IFT trains, it also failed to move directionally but diffused along the cilium instead[20] (see Supplementary Video 2, top left). This finding can now be explained by the direct role of the IFT-70(DYF-1) subunit in specific motor recruitment and activation as shown in this study. These results represent a functional reciprocity that is particularly notable given the complexity of the IFT machinery. Our work directly demonstrates the molecular mechanism by which an IFT motor is recruited to its IFT train, and is an example of the power of bottom-up approaches to delineate the underlying mechanisms of highly convoluted processes such as ciliogenesis.

Online content

Any Methods, including any statements of data availability and Nature Research reporting summaries, along with any additional references and Source Data files, are available in the online version of the paper at https://doi.org/10.1038/s41586-018-0105-3.

Received: 2 November 2017; Accepted: 23 March 2018;

1. Sharma, N., Berbari, N. F. & Yoder, B. K. Ciliary dysfunction in developmental abnormalities and diseases. *Curr. Top. Dev. Biol.* **85**, 371–427 (2008).
2. Pazour, G. J. & Rosenbaum, J. L. Intraflagellar transport and cilia-dependent diseases. *Trends Cell Biol.* **12**, 551–555 (2002).
3. Marshall, W. F. The cell biological basis of ciliary disease. *J. Cell Biol.* **180**, 17–21 (2008).
4. Goetz, S. C. & Anderson, K. V. The primary cilium: a signalling centre during vertebrate development. *Nat. Rev. Genet.* **11**, 331–344 (2010).
5. Pedersen, L. B. & Rosenbaum, J. L. Intraflagellar transport (IFT) role in ciliary assembly, resorption and signalling. *Curr. Top. Dev. Biol.* **85**, 23–61 (2008).
6. Kozminski, K. G., Johnson, K. A., Forscher, P. & Rosenbaum, J. L. A motility in the eukaryotic flagellum unrelated to flagellar beating. *Proc. Natl Acad. Sci. USA* **90**, 5519–5523 (1993).
7. Prevo, B., Scholey, J. M. & Peterman, E. J. G. Intraflagellar transport: mechanisms of motor action, cooperation, and cargo delivery. *FEBS J.* **284**, 2905–2931 (2017).
8. Avidor-Reiss, T. & Leroux, M. R. Shared and distinct mechanisms of compartmentalized and cytosolic ciliogenesis. *Curr. Biol.* **25**, R1143–R1150 (2015).
9. Scholey, J. M. Intraflagellar transport. *Annu. Rev. Cell Dev. Biol.* **19**, 423–443 (2003).
10. Rosenbaum, J. L. & Witman, G. B. Intraflagellar transport. *Nat. Rev. Mol. Cell Biol.* **3**, 813–825 (2002).
11. Taschner, M. & Lorentzen, E. The intraflagellar transport machinery. *Cold Spring Harb. Perspect. Biol.* **8**, a028092 (2016).
12. Snow, J. J. et al. Two anterograde intraflagellar transport motors cooperate to build sensory cilia on *C. elegans* neurons. *Nat. Cell Biol.* **6**, 1109–1113 (2004).
13. Mukhopadhyay, S. et al. Distinct IFT mechanisms contribute to the generation of ciliary structural diversity in *C. elegans*. *EMBO J.* **26**, 2966–2980 (2007).
14. Evans, J. E. et al. Functional modulation of IFT kinesins extends the sensory repertoire of ciliated neurons in *Caenorhabditis elegans*. *J. Cell Biol.* **172**, 663–669 (2006).
15. Jenkins, P. M. et al. Ciliary targeting of olfactory CNG channels requires the CNGB1b subunit and the kinesin-2 motor protein, KIF17. *Curr. Biol.* **16**, 1211–1216 (2006).
16. Leaf, A. & Von Zastrow, M. Dopamine receptors reveal an essential role of IFT-B, KIF17, and Rab23 in delivering specific receptors to primary cilia. *eLife* **4**, e06996 (2015).
17. Silverman, M. A. & Leroux, M. R. Intraflagellar transport and the generation of dynamic, structurally and functionally diverse cilia. *Trends Cell Biol.* **19**, 306–316 (2009).
18. Bae, Y. K. & Barr, M. M. Sensory roles of neuronal cilia: cilia development, morphogenesis, and function in *C. elegans*. *Front. Biosci.* **13**, 5959–5974 (2008).
19. Bae, Y. K. et al. General and cell-type specific mechanisms target TRPP2/PKD-2 to cilia. *Development* **133**, 3859–3870 (2006).
20. Ou, G., Blacque, O. E., Snow, J. J., Leroux, M. R. & Scholey, J. M. Functional coordination of intraflagellar transport motors. *Nature* **436**, 583–587 (2005).
21. Hao, L. et al. Intraflagellar transport delivers tubulin isotypes to sensory cilium middle and distal segments. *Nat. Cell Biol.* **13**, 790–798 (2011).

22. Ou, G. et al. Sensory ciliogenesis in *Caenorhabditis elegans*: assignment of IFT components into distinct modules based on transport and phenotypic profiles. *Mol. Biol. Cell* **18**, 1554–1569 (2007).
23. Burghoorn, J. et al. Mutation of the MAP kinase DYF-5 affects docking and undocking of kinesin-2 motors and reduces their speed in the cilia of *Caenorhabditis elegans*. *Proc. Natl Acad. Sci. USA* **104**, 7157–7162 (2007).
24. Masyukova, S. V. et al. A screen for modifiers of cilia phenotypes reveals novel MKS alleles and uncovers a specific genetic interaction between *osm-3* and *nphp-4*. *PLoS Genet.* **12**, e1005841 (2016).
25. Taschner, M. et al. Intraflagellar transport proteins 172, 80, 57, 54, 38, and 20 form a stable tubulin-binding IFT-B2 complex. *EMBO J.* **35**, 773–790 (2016).
26. Taschner, M., Bhogaraju, S. & Lorentzen, E. Architecture and function of IFT complex proteins in ciliogenesis. *Differentiation* **83**, S12–S22 (2012).
27. Piperno, G. & Mead, K. Transport of a novel complex in the cytoplasmic matrix of *Chlamydomonas* flagella. *Proc. Natl Acad. Sci. USA* **94**, 4457–4462 (1997).
28. Cole, D. G. et al. *Chlamydomonas* kinesin-II-dependent intraflagellar transport (IFT): IFT particles contain proteins required for ciliary assembly in *Caenorhabditis elegans* sensory neurons. *J. Cell Biol.* **141**, 993–1008 (1998).
29. Taschner, M., Bhogaraju, S., Vetter, M., Morawetz, M. & Lorentzen, E. Biochemical mapping of interactions within the intraflagellar transport (IFT) B core complex: IFT52 binds directly to four other IFT-B subunits. *J. Biol. Chem.* **286**, 26344–26352 (2011).
30. Imanishi, M., Endres, N. F., Gennerich, A. & Vale, R. D. Autoinhibition regulates the motility of the *C. elegans* intraflagellar transport motor OSM-3. *J. Cell Biol.* **174**, 931–937 (2006).
31. Prevo, B., Mangeol, P., Oswald, F., Scholey, J. M. & Peterman, E. J. Functional differentiation of cooperating kinesin-2 motors orchestrates cargo import and transport in *C. elegans* cilia. *Nat. Cell Biol.* **17**, 1536–1545 (2015).

Acknowledgements We thank G. Woehlke, E. J. G. Peterman, E. Lorentzen, M. Taschner and F. Müller-Planitz for discussions throughout this work, T.-H. Ho for technical assistance and A. Oberhofer and F. Müller-Planitz for critically reading the manuscript. This work was supported by European Research Council Grant 335623 (to Z.Ö.). We apologize to our colleagues whose work could not be cited owing to space limitations.

Reviewer information *Nature* thanks R. Vale and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions M.A.A.M. and Z.Ö. designed the experiments. M.A.A.M. performed the experiments and analysed the data. W.L.S. wrote all the customized MATLAB routines. Z.Ö. and M.A.A.M. wrote the manuscript.

Competing interests The authors declare no competing interests.

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0105-3.

Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0105-3.

Reprints and permissions information is available at http://www.nature.com/reprints.

Correspondence and requests for materials should be addressed to Z.Ö.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

No statistical methods were used to predetermine sample size. The experiments were not randomized and the investigators were not blinded to allocation during experiments and outcome assessment.

DNA and protein constructs. All DNA constructs used in this study were commercially synthesized (GenScript), they were based on sequence information that is available from the WormBase (http://www.wormbase.org) and cloned into the pFastBac1 vector according to the manufacturer's instructions (Thermo Fisher). In addition, the pFastBac1 vector was functionalized with Halo, SNAP, or GFP genes for C-terminal tagging of several constructs as described in the manuscript. Three different OSM-3 constructs were designed for this study (see Supplementary Information for details of DNA and protein sequences).

Protein expression, purification and fluorescent labelling. The Baculovirus Expression System (Thermo Fisher) was used to express all proteins in insect cells (Sf9) according to the manufacturer's instructions.

For protein purification, Sf9 cells at a concentration of 2×10^6 cells ml^{-1} were infected with the corresponding viruses. After 65 h of incubation at 28 °C, the cells were collected by centrifugation for 15 min at 2,600*g*. Cell pellets were lysed in PIPES buffer (50 mM PIPES, pH 6.9, 300 mM potassium acetate, 1 mM $MgCl_2$, 1 mM dithiothreitol (DTT), 0.5% Triton X-100, 10% glycerol, 0.1 mM ATP, Complete Protease Inhibitor Cocktail (Roche)). Lysed cells were pelleted by centrifugation for 10 min at 40,000*g*. The supernatant was incubated with 60 μl of Anti-Flag M2 Affinity Gel (Sigma) for 90 min. The Flag-resin was washed three times with 1 ml of wash buffer 1 (80 mM PIPES, pH 6.9, 500 mM potassium acetate, 1 mM $MgCl_2$, 1 mM DTT, 5 μM ATP, 10% glycerol, 0.1% Tween-20, 1 mM EGTA) and three times with 1 ml of wash buffer 2 (80 mM PIPES, pH 6.9, 200 mM potassium acetate, 1 mM $MgCl_2$, 1 mM DTT, 0.1 mM ATP, 10% glycerol, 0.1% Tween-20, 1 mM EGTA). The Flag resin was incubated in a rotator for 30 min at room temperature with 150 μl wash buffer 2 containing either 1 mM HaloTag Alexa Fluor660 ligand (for labelling OSM-3(G444E)–Halo) or 1 mM SNAP-Surface Alexa Fluor 647 or 488 (for labelling OSM-3–SNAP and the corresponding IFT-B subunits). The Flag resin was washed three times with 1 ml of wash buffer 2 and eluted with 70 μl of elution buffer (wash buffer 2 containing 0.5 mg ml^{-1} of 1 × Flag Peptide (Sigma)) for 60 min at 4 °C[32,33].

SEC–MALS. SEC–MALS analyses were used to determine the absolute molar masses of the protein complexes. The TSKgel G4000SW$_{XL}$ (Tosoh Biosciences) and the Superose 6 Increase 10/300 GL columns (GE Healthcare) were calibrated with at least two column volumes of the gel filtration buffer (25 mM PIPES, pH 7.0, 200 mM NaCl, 1 mM $MgCl_2$, 1 mM EGTA, 1 mM DTT). The columns were inline with a variable UV-absorbance detector (Agilent 1260 Infinity series) and a DAWN8 + MALS detector (Wyatt Technology). Molar masses were calculated with ASTRA 6 software (Wyatt Technology) with the d*n*/d*c* value set to 0.185 ml g^{-1}. Bovine serum albumin (BSA) was used as a calibration standard.

MST assay. OSM-3(G444E)–Halo interaction with either the QCC or the QCC lacking IFT70(DYF-1) was measured using MST with a Monolith NT.115 (NanoTemper Technologies GmbH). The fluorophore-labelled OSM-3(G444E)–Halo at 10 nM was titrated with the respective (unlabelled) binding partner, the concentration of which varied between 0.2 nM and 2 μM in the MST buffer (50 mM Tris-HCl (pH 7.6), 150 mM NaCl, 1 mM $MgCl_2$, 0.1% Tween 20, 1 mM EGTA, 1 mM DTT, 10% glycerol). Standard coated capillaries were used in all measurements. K_d values were determined using NanoTemper Analysis software.

Photobleaching experiments and single-molecule assays. Photobleaching experiments and single-molecule assays were conducted as described previously[32,33]. In brief, after purifying the proteins as described above, each motor was mixed with its corresponding IFT-B proteins in an equimolar ratio and incubated overnight at 4 °C in a rotator. To track the movement of the motor alone or in the presence of IFT-B proteins, microtubules were attached to the surface of a flow chamber (coated with 1 mg ml^{-1} biotinylated BSA and 1 mg ml^{-1} streptavidin (Sigma)) and the fluorescently labelled proteins were diluted to the desired concentration in BRB80 motility buffer (80 mM PIPES, pH 6.9, 10 mM ATP, 0.145 mg ml^{-1} glucose oxidase (Sigma), 0.0485 mg ml^{-1} catalase (Sigma), 20% glucose) and flowed into the chamber. Movement of the fluorescent signals was recorded with a cycle time of 235 ms with an objective-type Leica DMI6000 B TIRF microscope (Leica), equipped with a plan objective lens (100×, numerical aperture (NA) 1.47 oil), and a back-illuminated Andor U897 EMCCD camera (Andor). Excitation was achieved with the help of diode lasers at 488 and 635 nm wavelength, and frames were recorded and analysed with AF 6000 software (Leica). The velocities and run lengths were analysed with custom-written programs using MATLAB software (Mathworks). Spots were selected automatically according to their brightness compared with the mean brightness in each frame. The position of the spots was determined with subpixel accuracy using a radial centre approach. Runs were considered processive with a minimal run length of 1 μm. The run length data were fit to a truncated ($\chi_0 = 1$ μm) single-exponential distribution. Parts of the distance-over-time data were considered for speed calculation in a linear fit of at least six frames that resulted in a $r^2 > 95\%$. Landing events were counted from the data obtained using the tracking algorithm described above. Events that started after the first frame and showed a displacement from the binding position of at least 2 μm while associated with the microtubule were counted (n). As no processive movement was observed for the OSM-3–SNAP, runs were not selected for unidirectionality. Images of the microtubule positions were obtained from maximum intensity projections of the single molecule movies and their length (l) were measured using ImageJ. Considering the total length of a movie (t), the landing rate for a movie was calculated as $e = n/l/t$. This value was then corrected for relative dilution (c) of the motor over the motor plus adaptor) ($e_{corr} = e/c$). The mean of this value over three independent purifications was calculated alongside the s.d. and reported. For the colocalized movements, runs were compared pairwise from the two channels. Weighted penalties resulted from the mean distances of tracked positions (pixel, factor 1/3) and the difference of the starting time (frames, factor 1). Runs with a penalty below ten were considered and parameters are averages of single runs.

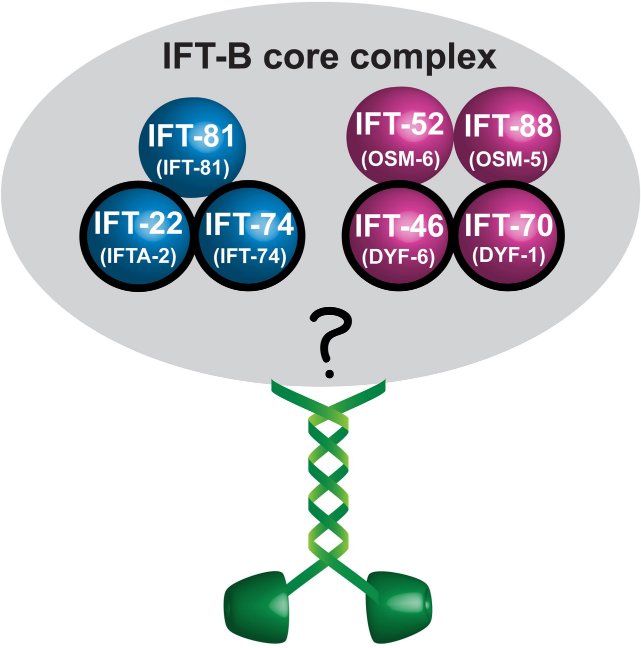
Colocalization assays. Proteins were mixed in equimolar concentrations, incubated overnight at 4 °C and flowed into chambers as described above for single-molecule assays. Fluorescence was detected using TIRF illumination. The colocalized images were analysed using custom-written routines in MATLAB software (MathWorks).

Reporting summary. Further information on experimental design is available in the Nature Research Reporting Summary linked to this paper.

Code availability. Custom-written codes used in this study are available from the corresponding author upon reasonable request.

Data availability. All data that support the findings of this study, including uncropped SDS–PAGE analyses of purified complexes, the detailed sequences of the motor constructs used in the study, are contained within the paper and its Supplementary Information or are available from the corresponding author upon reasonable request. Source Data for Figs. 2, 3 and Extended Data Figs. 4, 5, 7 are available with the online version of the paper.

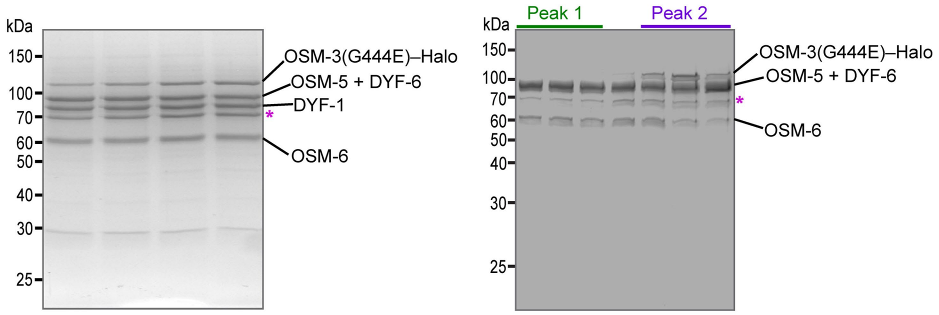
32. Oberhofer, A. et al. Myosin Va's adaptor protein melanophilin enforces track selection on the microtubule and actin networks in vitro. *Proc. Natl Acad. Sci. USA* **114**, E4714–E4723 (2017).
33. Stepp, W. L., Merck, G., Mueller-Planitz, F. & Ökten, Z. Kinesin-2 motors adapt their stepping behavior for processive transport on axonemes and microtubules. *EMBO Rep.* **18**, 1947–1956 (2017).

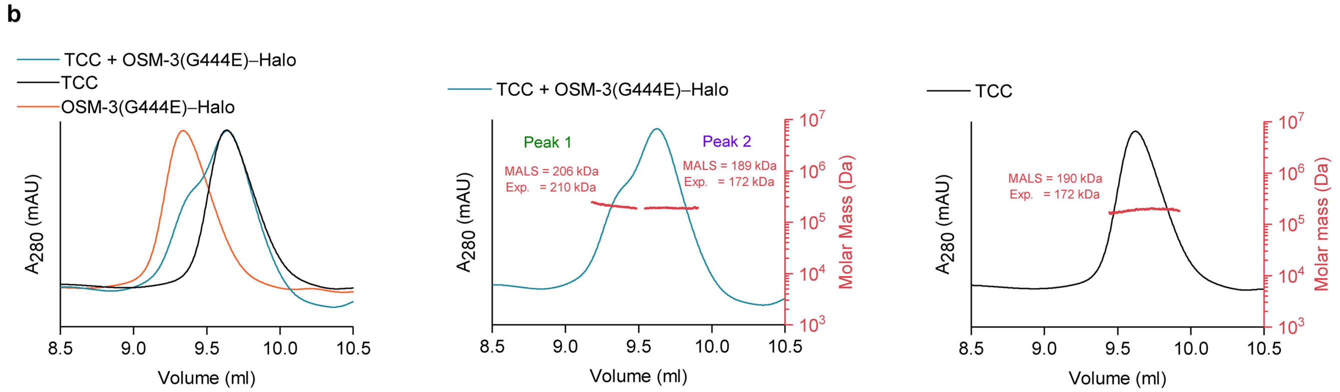


Extended Data Fig. 1 | Overview of recombinant constructs and their proposed assembly in vitro. Schematic of the presumptive IFT-B core complex from *C. elegans* (corresponding nomenclature of the subunits in *C. elegans* is shown in brackets). Subunits known to autonomously form sub-complexes in *C. reinhardtii* are colour-coded. The OSM-3 motor is shown in green. Subunits of the IFT-B core complex that are proposed to interfere with OSM-3 function in vivo are highlighted with black circles.

a

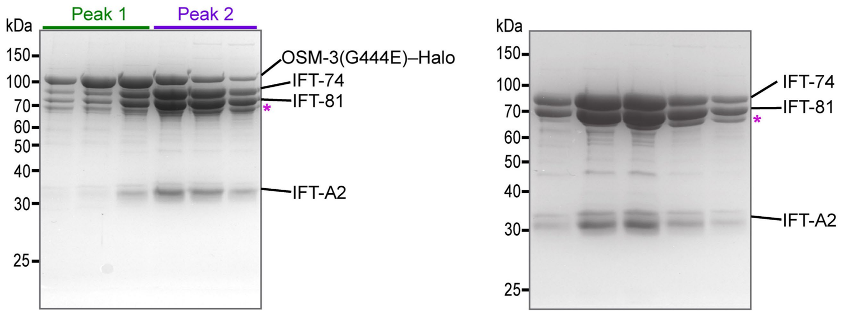
Protein	Calculated molar mass (kDa)
OSM-6	55
DYF-6	55
OSM-5	92
DYF-1	73
OSM-3(G444E)–Halo	210
Total	485



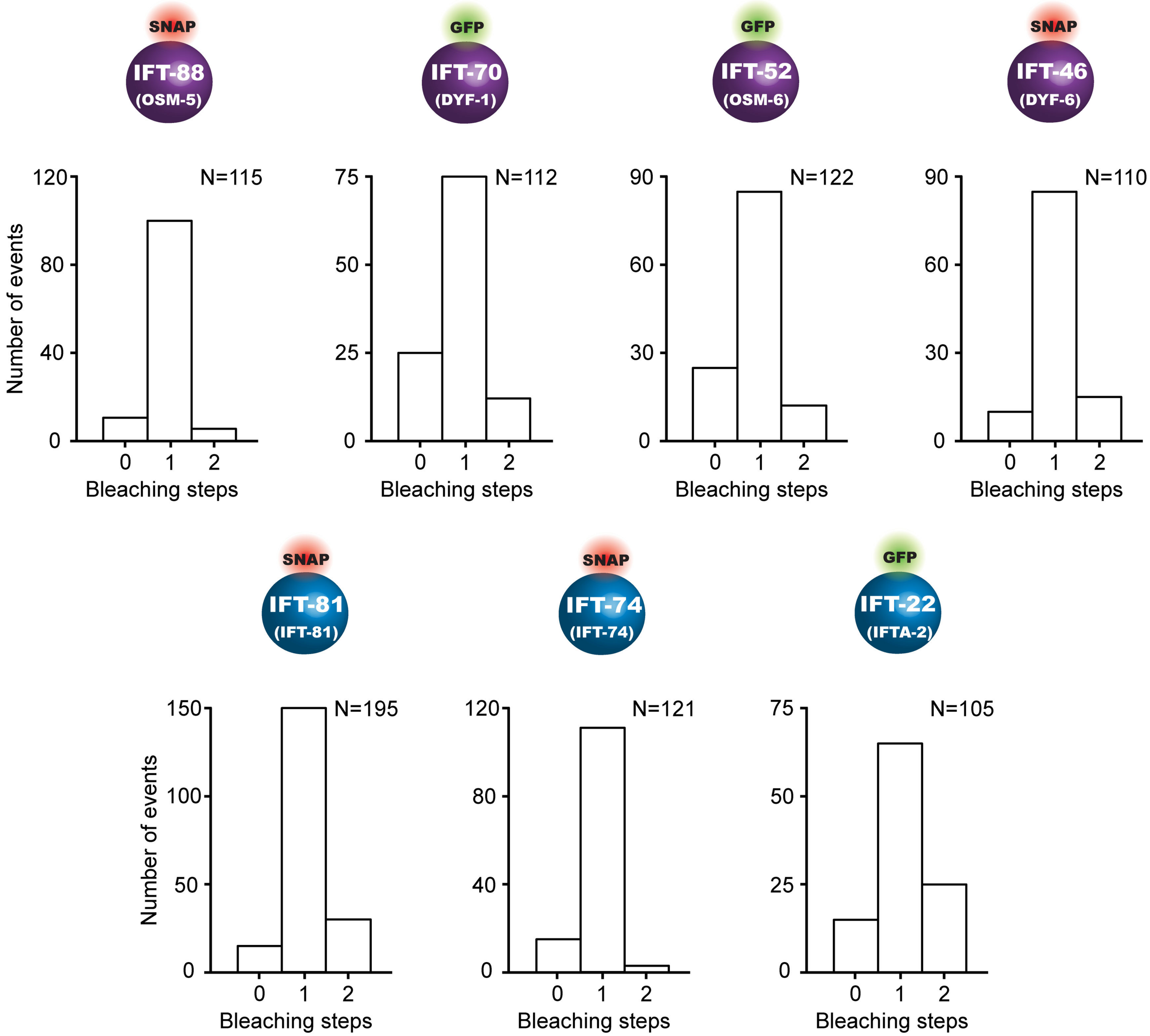


c

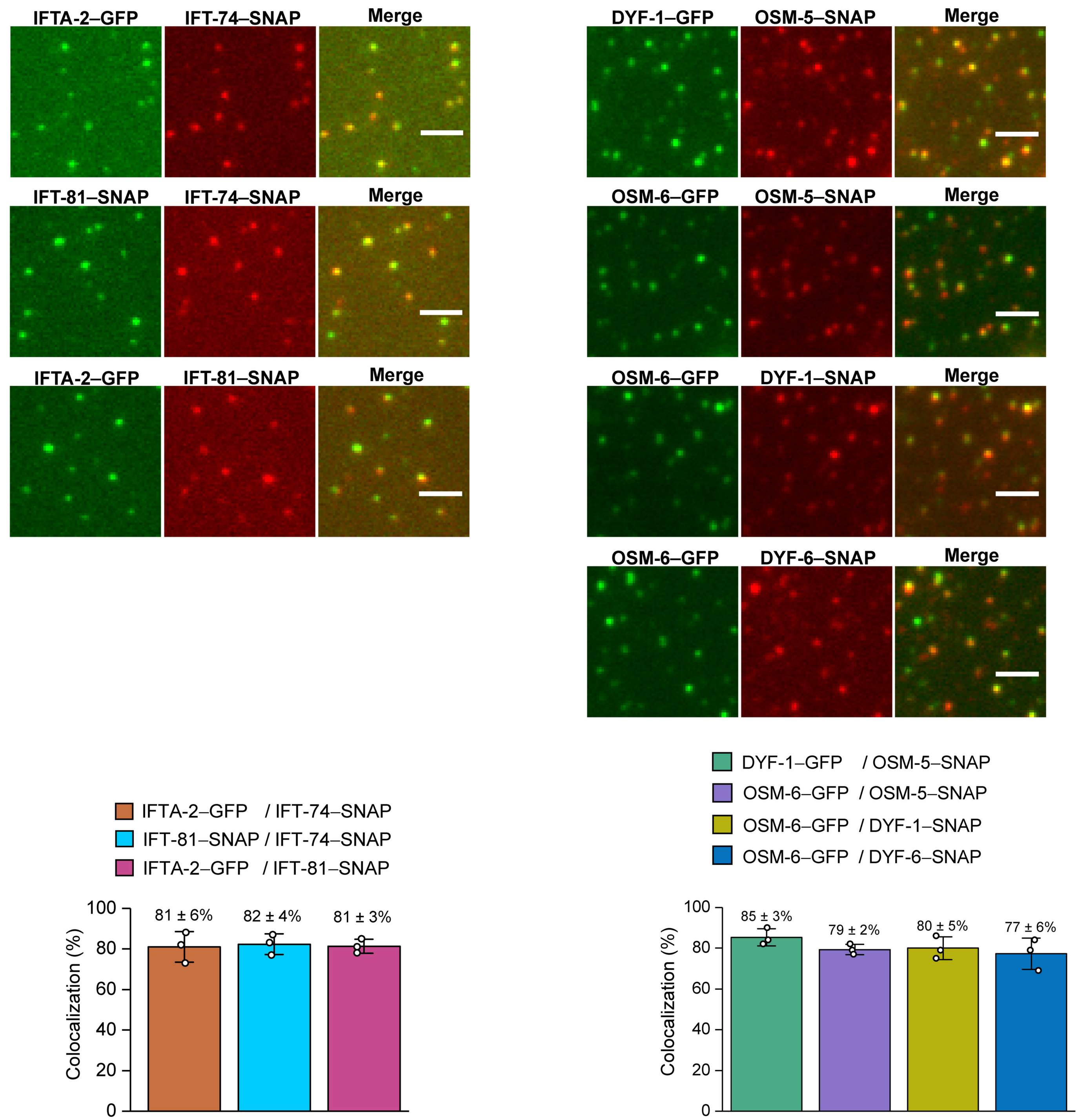
Protein	Calculated molar mass (kDa)
IFT-74	73
IFT-81	70
IFT-A2	29
OSM-3(G444E)–Halo	210
Total	382



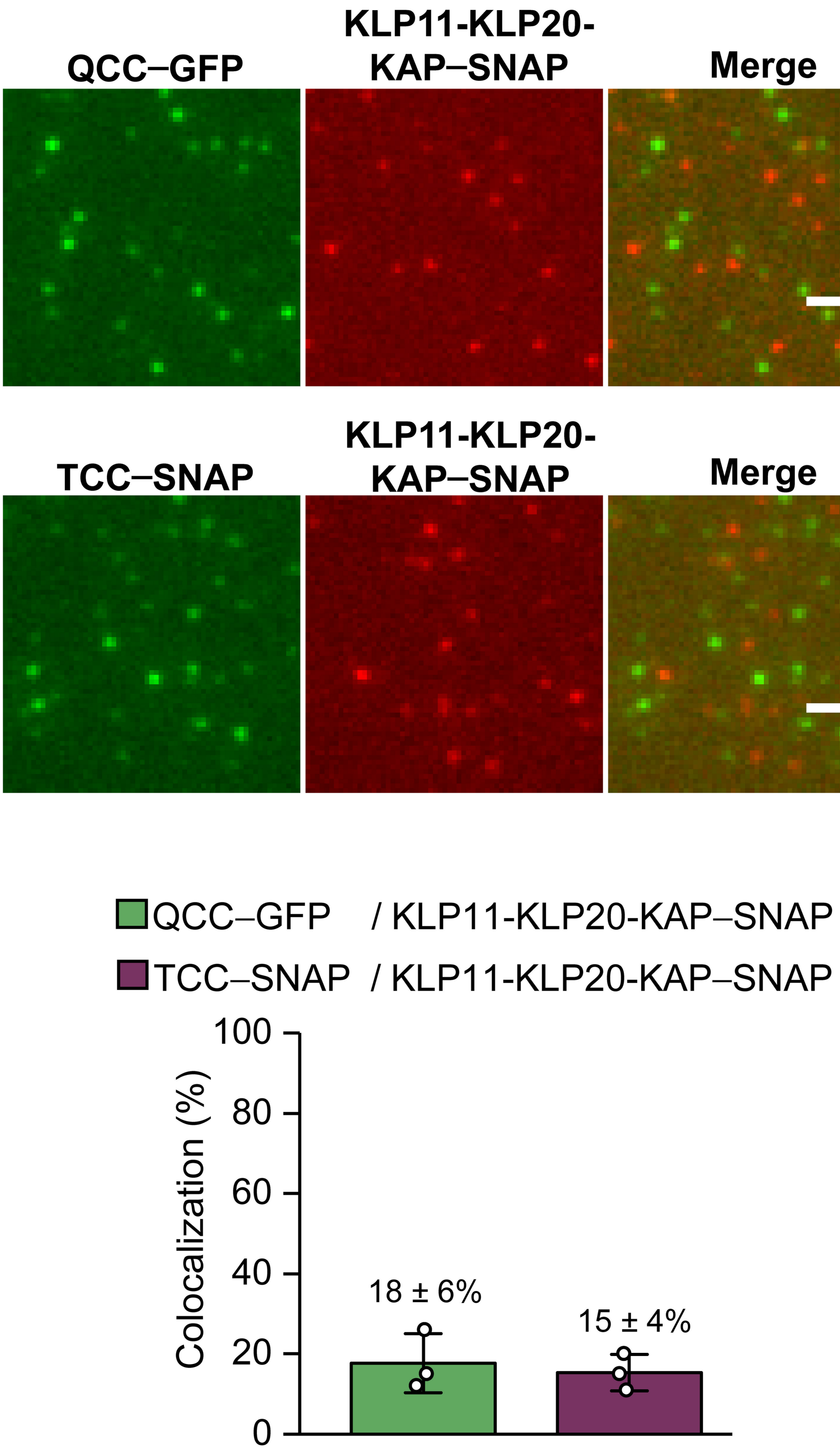
Extended Data Fig. 2 | Interaction of OSM-3(G444E)–Halo with QCC and TCC. a, Calculated molar mass of QCC subunits and the OSM-3(G444E)–Halo motor along with their expected sum (left). The SDS–PAGE analyses of the elution peaks in Fig. 1 (middle and right) show that OSM-3(G444E)–Halo co-elutes with the complex in the presence of the IFT-70(DYF-1) subunit (middle) but does not co-elute with the complex in the absence (right) of the IFT-70(DYF-1) subunit. **b**, Overlay of the elution profiles of the TCC with and without the OSM-3(G444E)–Halo motor and of the OSM-3(G444E)–Halo motor alone (left). Note that the left shoulder of the TCC + OSM-3(G444E)–Halo complex overlaps with the elution profile of the OSM-3(G444E)–Halo motor, and the right shoulder with the TCC. Consistently, the molar masses determined for the TCC + OSM-3(G444E)–Halo under peak 1 correspond to the OSM-3(G444E)–Halo motor and peak 2 to the TCC, respectively (middle versus right). **c**, Left, the calculated molar mass of the TCC subunits and the OSM-3(G444E)–Halo along with their expected sum. Middle and right, the SDS–PAGE analyses of the elution peaks shown in **b**. Data are representative of three independent experiments. The identities of all subunits were confirmed by liquid chromatography–tandem mass spectrometry (LC–MS/MS) analysis. Asterisks in **a** and **c** indicate the location of Hsp70 protein.



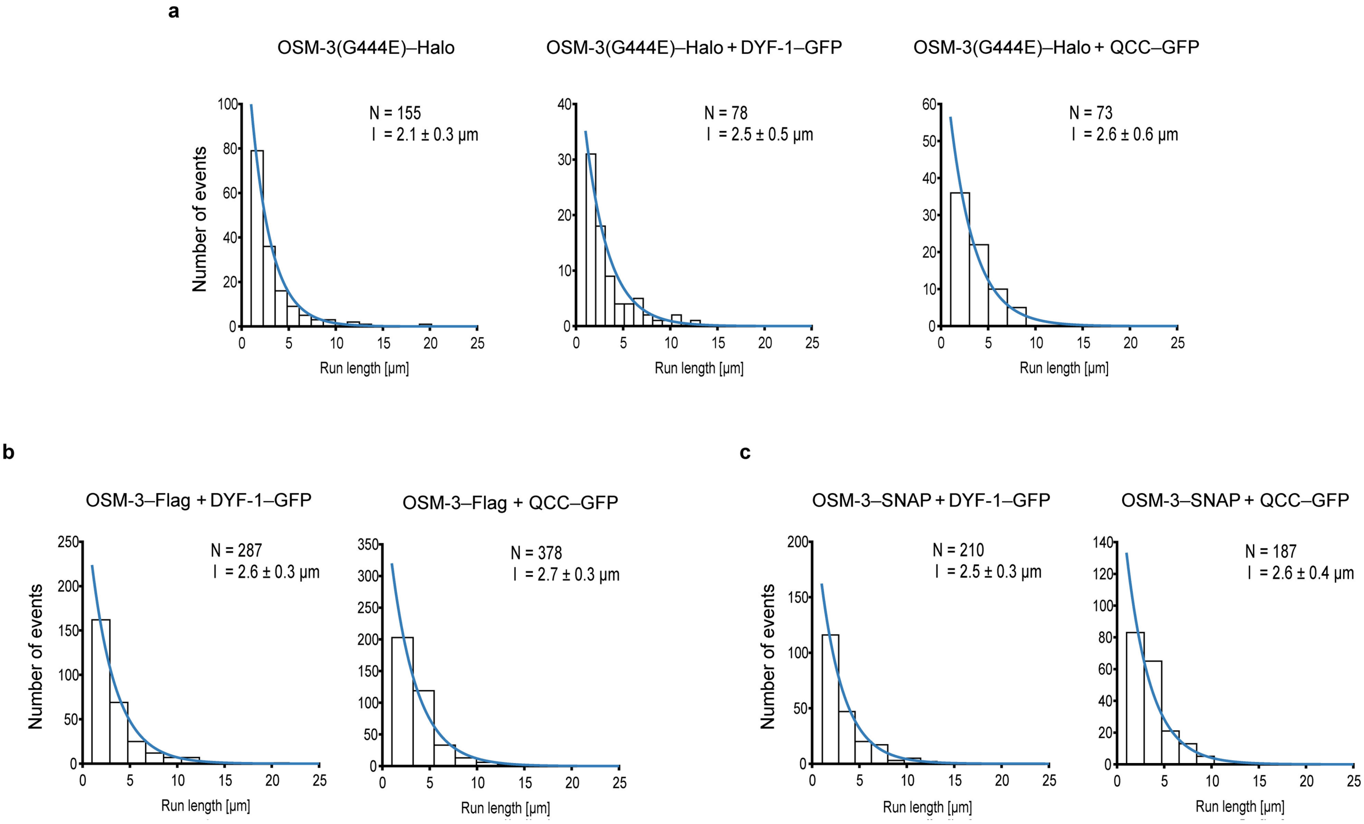
Extended Data Fig. 3 | Overview of the IFT-B subunits functionalized with C-terminal tags and their photobleaching properties. The subunits were functionalized either with a GFP or SNAP tag for fluorescence labelling. All subunits displayed mostly single-step photobleaching consistent with non-aggregated, single subunits after functionalization. *N*, the number of events obtained from three different slides in three independent experiments.



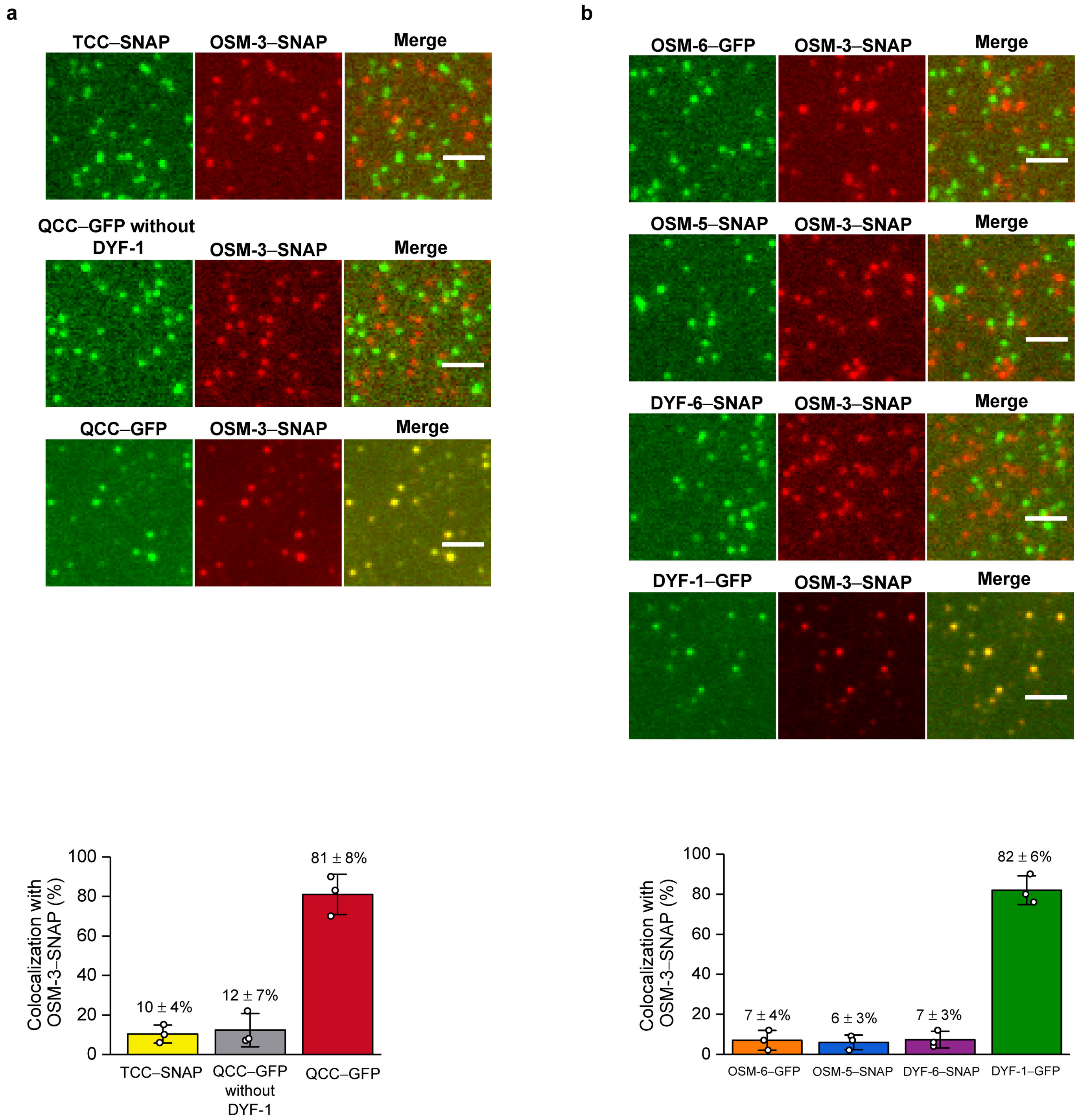
Extended Data Fig. 4 | Pairwise colocalization efficiency of IFT-B subunits. Pairs of differentially labelled subunits of the IFT-B subcomplexes were incubated and analysed for their colocalization efficiency. The columns (bottom) represent the percentage of colocalized spots in the corresponding colocalized images (top). All assayed combinations of the labelled subunits displayed significant colocalization efficiencies demonstrating that C-terminal functionalization of the subunits does not interfere with their complex formation capabilities. Data are mean ± s.d. from three independent experiments. Scale bars, 3 μm. *C. elegans* nomenclature is used in this figure owing to space limitations.



Extended Data Fig. 5 | Colocalization of the heterotrimeric KLP11–KLP20–KAP motor with the IFT-B complex. Neither QCC (top) nor TCC (bottom) of the IFT-B complex displayed efficient colocalization with the KLP11–KLP20–KAP motor. The IFT-81 subunit of TCC was fluorescently labelled with a SNAP tag and the IFT-52(OSM-6) subunit of QCC was GFP tagged. Data are mean ± s.d. from three independent experiments. Scale bars, 3 µm.



Extended Data Fig. 6 | The IFT-70(DYF-1)-dependent activation does not alter the processivity of the OSM-3 motor. a–c, OSM-3(G444E)–Halo (**a**), and the IFT-70(DYF-1)-activated OSM-3–Flag (**b**), and OSM-3–SNAP (**c**) motors display similar processivity that is independent of the presence of the IFT-70(DYF-1) subunit and the QCC. *N*, the number of events obtained from three different flow chambers in three independent experiments. Run length was fit to a single exponential ± confidence interval.



Extended Data Fig. 7 | OSM-3–SNAP motor containing the wild type stalk colocalized with QCC in an IFT-70(DYF-1)-dependent manner. **a**, Neither TCC nor QCC lacking DYF-1 efficiently colocalize with the OSM-3–SNAP motor. However in the presence of the IFT-70(DYF-1) subunit, the QCC efficiently colocalizes with OSM-3–SNAP (81 ± 8%). **b**, Consistently, OSM-3–SNAP showed robust colocalization (82 ± 6%) with the IFT-70(DYF-1) subunit but not with IFT-52(OSM-6), IFT-88(OSM-5) or IFT-46(DYF-6) subunits. IFT-81 from TCC was fluorescently labelled with a SNAP tag and IFT-52(OSM-6) from QCC and QCC lacking DYF-1 were GFP-tagged. Data are mean ± s.d. of three independent experiments. Scale bars, 3 µm. *C. elegans* nomenclature is used in the figure owing to space limitations.

https://doi.org/10.1038/s41586-018-0117-z

The onset of star formation 250 million years after the Big Bang

Takuya Hashimoto[1,2]*, Nicolas Laporte[3,4], Ken Mawatari[1], Richard S. Ellis[3], Akio K. Inoue[1], Erik Zackrisson[5], Guido Roberts-Borsani[3], Wei Zheng[6], Yoichi Tamura[7], Franz E. Bauer[8,9,10], Thomas Fletcher[3], Yuichi Harikane[11,12], Bunyo Hatsukade[13], Natsuki H. Hayatsu[12,14], Yuichi Matsuda[2,15], Hiroshi Matsuo[2,15], Takashi Okamoto[16], Masami Ouchi[11,17], Roser Pelló[4], Claes-Erik Rydberg[18], Ikkoh Shimizu[19], Yoshiaki Taniguchi[20], Hideki Umehata[13,20,21] & Naoki Yoshida[12,17]

A fundamental quest of modern astronomy is to locate the earliest galaxies and study how they influenced the intergalactic medium a few hundred million years after the Big Bang[1–3]. The abundance of star-forming galaxies is known to decline[4,5] from redshifts of about 6 to 10, but a key question is the extent of star formation at even earlier times, corresponding to the period when the first galaxies might have emerged. Here we report spectroscopic observations of MACS1149-JD1[6], a gravitationally lensed galaxy observed when the Universe was less than four per cent of its present age. We detect an emission line of doubly ionized oxygen at a redshift of 9.1096 ± 0.0006, with an uncertainty of one standard deviation. This precisely determined redshift indicates that the red rest-frame optical colour arises from a dominant stellar component that formed about 250 million years after the Big Bang, corresponding to a redshift of about 15. Our results indicate that it may be possible to detect such early episodes of star formation in similar galaxies with future telescopes.

Between March 2016 and April 2017 we performed observations of MACS1149-JD1 at the Atacama Large Millimeter/submillimeter Array (ALMA), targeting the far-infrared oxygen line, [O III] at a wavelength of 88 μm, and dust continuum emission over a broad wavelength range consistent with its photometric redshift range (9.0–9.8)[4,7–9]. The [O III] line is clearly detected at a redshift of $z = 9.1096 \pm 0.0006$ with a peak intensity of 129.8 ± 17.5 mJy km s^{-1} per beam, corresponding to the significance level of 7.4σ (Fig. 1, Table 1). (1σ error values reported in this paper denote the 1σ root-mean-square or standard deviation unless otherwise specified.) The spatial location is coincident with the rest-frame ultraviolet continuum emission detected by the Hubble Space telescope (HST), indicating that the [O III] line arises from a star-forming galaxy. The line has a luminosity of $(7.4 \pm 1.6)(10/\mu) \times 10^{7} L_{\odot}$, where $L_{\odot}$ is the luminosity of the Sun and for the lensing magnification we adopt a fiducial value of $\mu = 10$ (see Methods). Its full-width at half-maximum (FWHM) is 154 ± 39 km s^{-1}, representative of that seen in other high-z galaxies[10]. The [O III] emission is spatially resolved,

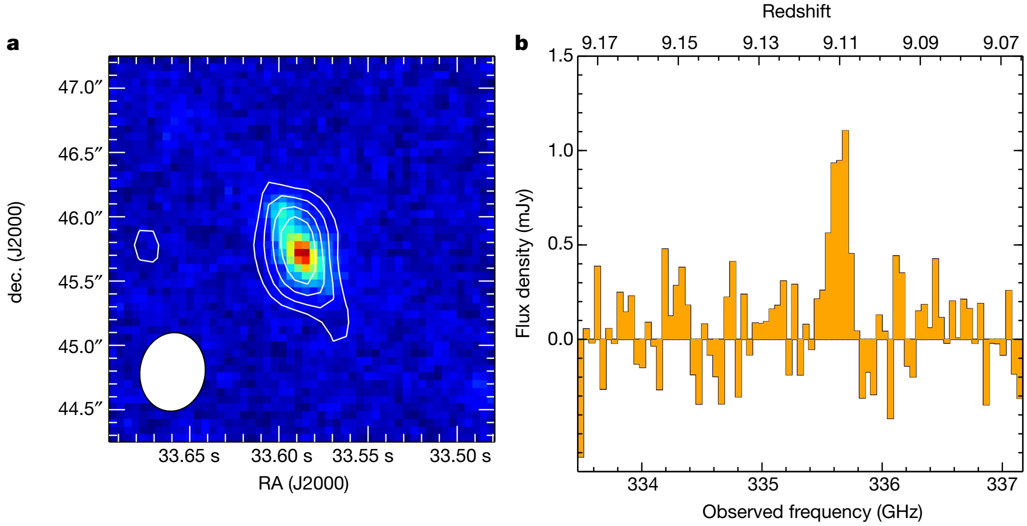


Fig. 1 | ALMA [O III] contours and spectrum of MACS1149-JD1. **a**, Magnification of an HST image (F160W), with the ALMA [O III] contours overlaid. Contours are drawn at -3σ (dashed line), 3σ, 4σ, 5σ and 6σ, where $\sigma = 17.5$ mJy km s^{-1} per beam. The ellipse at the lower left corner indicates the synthesized beam size of ALMA. RA is the right ascension and dec. denotes the declination, both expressed in J2000 coordinates. **b**, The ALMA [O III] 88-μm spectrum in frequency space, obtained with a resolution of about 42 km s^{-1}.

---

[1] Department of Environmental Science and Technology, Faculty of Design Technology, Osaka Sangyo University, Osaka, Japan. [2] National Astronomical Observatory of Japan, Tokyo, Japan. [3] Department of Physics and Astronomy, University College London, London, UK. [4] IRAP, Université de Toulouse, CNRS, UPS, CNES, Toulouse, France. [5] Department of Physics and Astronomy, Uppsala University, Uppsala, Sweden. [6] Department of Physics & Astronomy, Johns Hopkins University, Baltimore, MD, USA. [7] Division of Particle and Astrophysical Science, Graduate School of Science, Nagoya University, Nagoya, Japan. [8] Instituto de Astrofísica and Centro de Astroingeniería, Facultad de Física, Pontificia Universidad Católica de Chile, Santiago, Chile. [9] Millennium Institute of Astrophysics (MAS), Santiago, Chile. [10] Space Science Institute, Boulder, CO, USA. [11] Institute for Cosmic Ray Research, The University of Tokyo, Chiba, Japan. [12] Department of Physics, Graduate School of Science, The University of Tokyo, Tokyo, Japan. [13] Institute of Astronomy, The University of Tokyo, Tokyo, Japan. [14] European Southern Observatory, Garching bei München, Germany. [15] Department of Astronomical Science, School of Physical Sciences, The Graduate University for Advanced Studies (SOKENDAI), Tokyo, Japan. [16] Department of Cosmosciences, Graduates School of Science, Hokakido University, Sapporo, Japan. [17] Kavli Institute for the Physics and Mathematics of the Universe (WPI), Todai Institutes for Advanced Study, The University of Tokyo, Chiba, Japan. [18] Universität Heidelberg, Zentrum für Astronomie, Institut für Theoretische Astrophysik, Heidelberg, Germany. [19] Theoretical Astrophysics, Department of Earth and Space Science, Osaka University, Osaka, Japan. [20] The Open University of Japan, Chiba, Japan. [21] The Institute of Physical and Chemical Research (RIKEN), Saitama, Japan. *e-mail: thashimoto@est.osaka-sandai.ac.jp

Table 1 | Properties of MACS1149-JD1

Parameters	Values
RA	11 h 49 min 33.58 s
dec.	+22° 24′ 45.7″
Redshift, $z_{[\mathrm{O\,III}]}$	9.1096 ± 0.0006
[O III] line width, $\mathrm{FWHM}_{[\mathrm{O\,III}]}$ (km s^{-1})	154 ± 39
[O III] luminosity, $L_{[\mathrm{O\,III}]}$ ($\times 10^7 L_\odot$)	$(7.4 \pm 1.6) \times (10/\mu)$
90-μm continuum flux (μJy per beam)	$< 5.3 \times (10/\mu)$ (3σ)
Star formation rate ($M_\odot$ yr^{-1})	$4.2^{+0.8}_{-1.1} \times (10/\mu)$
Stellar mass ($\times 10^9 M_\odot$)	$1.08^{+0.53}_{-0.18} \times (10/\mu)$
Dust mass[a] ($\times 10^5 M_\odot$)	$< 5.2 \times (10/\mu)$ (3σ)
Dynamical mass[b] ($\times 10^9 M_\odot$)	$(4 \pm 3) \times (10/\mu)^{0.5}$
Redshift, $z_{\mathrm{Ly}\alpha}$	9.0944 ± 0.0019
Velocity offset $\Delta v_{\mathrm{Ly}\alpha}$ (km s^{-1})	−450 ± 60
Lyα line width $\mathrm{FWHM}_{\mathrm{Ly}\alpha}$ (km s^{-1})	144 ± 56
Lyα luminosity $L_{\mathrm{Ly}\alpha}$ ($\times 10^7 L_\odot$)	$(12.4 \pm 3.2) \times (10/\mu)$

μ, Lensing magnification; RA, right ascension; dec., declination.

[a] We assume a dust temperature of $T_\mathrm{d} = 40$ K, a spectral index of $\beta_\mathrm{d} = 1.5$ and a dust-emitting region of a single beam size.

[b] This value is derived under the assumption that lensing effects are equal for the major and minor axes.

and its deconvolved size is (0.82″ ± 0.25″) × (0.30″ ± 0.14″). The corresponding intrinsic size is $(3.7 \pm 1.1)\mu^{-0.5}$ kpc × $(1.4 \pm 0.9)\mu^{-0.5}$ kpc, assuming that lensing effects are equal for the major and minor axes. Assuming that MACS1149-JD1 is a dispersion-dominated system, we derive a dynamical mass of $(4 \pm 3) \times (10/\mu)^{0.5} \times 10^9 M_\odot$, where $M_\odot$ is the mass of the Sun.

No dust continuum is detected above a 3σ upper limit of $S_{\nu,90\mu\mathrm{m}} < 5.3 \times (10/\mu)$ μJy per beam, where $S_{\nu,90\mu\mathrm{m}}$ corresponds to the flux density at a rest-frame wavelength of 90 μm. For a dust temperature of $T_\mathrm{d} = 40$ K and emissivity index $\beta_\mathrm{d} = 1.5$, the total infrared luminosity is smaller than $7.7(10/\mu) \times 10^9 L_\odot$ after correcting for the contribution of the cosmic microwave background[11]. Assuming a representative dust mass absorption coefficient[12], we derive a 3σ upper limit on a dust mass of $5.3(10/\mu) \times 10^5 M_\odot$.

The spectral energy distribution (SED) of MACS1149-JD1 has a prominent excess signal[6] in the Spitzer Infrared Array Camera (IRAC) channel-2 band at 4.5 μm. Previous spectroscopic studies on sources exhibiting such 'IRAC excess'[13–15] have claimed that such an excess probably arises from intense emission in the [O III] 5,007-Å line. However, the origin of the IRAC excess of MACS1149-JD1 remains unclear owing to the inaccuracy of the photometrically estimated redshift[6,7,16]. Our precisely determined redshift of $z = 9.1$ rules out the [O III] 5,007-Å line as the source of the excess because the wavelength of this line has < 1% transmission in the IRAC channel-2 bandpass. Given recent evidence that the [O III] line at 88 μm may not give an accurate indication of the systemic velocity of distant star-forming galaxies[17], we also present tentative evidence for Lyman α (Lyα) emission from MACS1149-JD1, obtained with the X-shooter[18] spectrograph at the European Southern Observatory's Very Large Telescope using 6.5 h of data taken between February and April 2017. A 4σ-significance detection at a wavelength of 12,271.51 Å, corresponding to Lyα at $z = 9.0942 \pm 0.0019$, provides a further valuable constraint on the source redshift.

Our detailed analysis (see Methods) shows that strong Hβ and [O III] 4,959 Å lines in IRAC channel 2 cannot reproduce the excess within the likely redshift range of the source because, for the required young ages, the IRAC channel-1 flux would also be considerably boosted by strong nebular continuum emission and [O II] 3,727-Å line emission. Such young models are further disfavoured by additional constraints from the measured ultraviolet continuum slope and the ALMA upper limit on the dust attenuation. Therefore, we conclude that the IRAC excess can only arise from the stellar continuum around the Balmer break at about 4,000 Å.

Although MACS1149-JD1 is elongated by the effects of gravitational lensing, there is some evidence of two components in the HST image (Fig. 1), so the IRAC excess at 4.5 μm may originate from a separate source. The peak of the ALMA [O III] emission is coincident with the brighter component (Fig. 1), whereas the less-well located Lyα emission and IRAC flux could arise in either component. However, the wide wavelength coverage of the SED enables us to reject the possibility that a low redshift interloper contributes to the 4.5-μm excess (see Methods for detailed analyses).

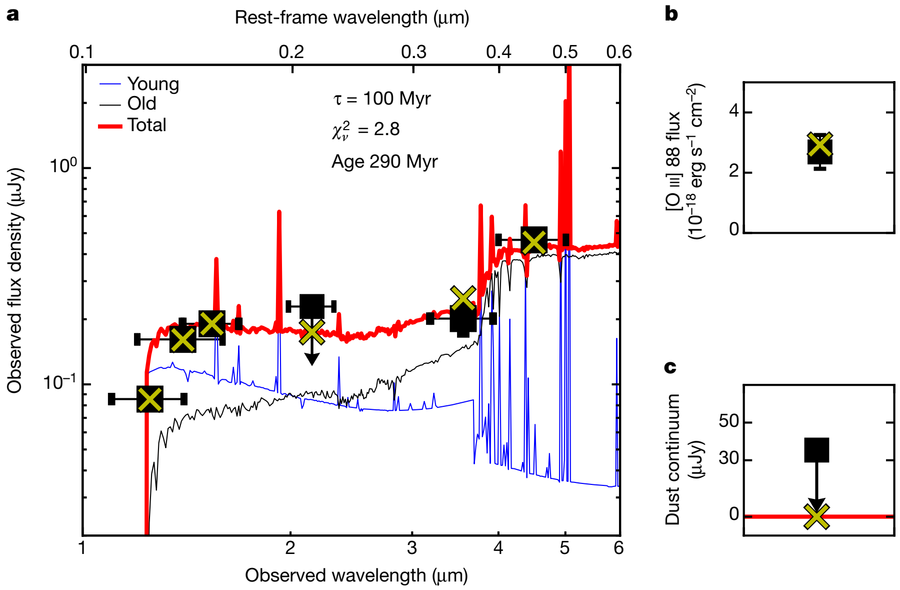


Fig. 2 | The SED of MACS1149-JD1. The [O III] flux provides a constraint on the star formation rate at the epoch of observation, while the SED gives valuable evidence of its earlier star formation history. The model shown involves a burst of star formation of duration τ = 100 Myr within the interval $12 < z < 15$ and reproduces the required excess flux at 4.5 μm. A much younger component reproduces the strength of [O III] emission observed at $z \approx 9$. **a**, Black squares show (from left to right) F125W, F140W and F160W data from HST[7], a 2σ upper limit for the $\mathrm{K_s}$ band from VLT/HAWK-I[16], and 3.6-μm and 4.5-μm fluxes from Spitzer/IRAC[7]. The horizontal and vertical error bars show the wavelength range of the filters and 1σ measurement uncertainties, respectively. The red solid line indicates the SED model and the corresponding magnitudes are shown by yellow crosses. Blue and black lines represent the contributions from the young and old component, respectively. **b**, The black square is the observed [O III] emission line flux and its 1σ uncertainty, while the yellow cross indicates the model prediction. **c**, The black square shows the 2σ upper limit for the dust continuum flux density, and the yellow cross indicates the model prediction.

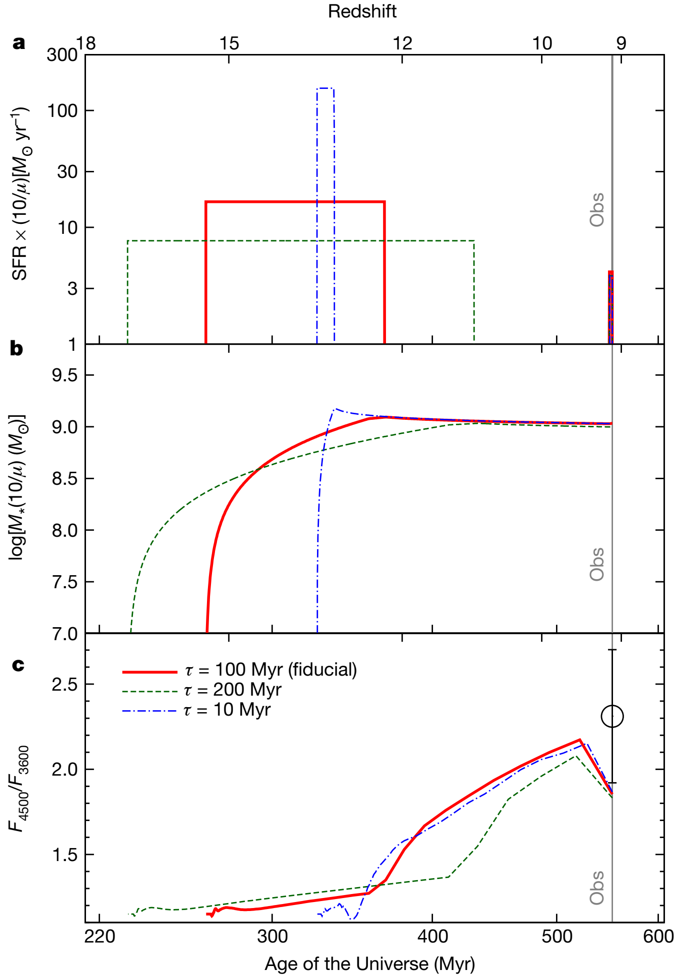


Fig. 3 | Demonstration of how a dominant phase of early star formation is necessary to reproduce the SED of MACS1149-JD1. a–c, We show the SFR (**a**), stellar mass assembly (**b**) and Balmer break (**c**) as a function of redshift for three possible models. F_{4500} and F_{3600} denote the flux densities at the rest-frame wavelengths of 4,500 Å and 3,600 Å, respectively. In each panel, the blue dot-dashed, green dashed and red solid lines correspond to the models with star formation bursts of duration 10 Myr, 100 Myr and 200 Myr, respectively. Each model is capable of reproducing the Balmer break observed at $z = 9.1$ (shown by the black data point). The full SED fits shown in Fig. 2 correspond to our fiducial model, which assumes a burst with a duration of 100 Myr. The bulk of the stellar mass M_* in MACS1149-JD1, observed at $z = 9.1$, was produced within a short period corresponding to the redshift interval $12 < z < 16$. Further details of these models and other assumptions are provided in Methods.

Independently of the SED, the ALMA [O III] emission line flux provides a valuable measure of the star formation rate (SFR) at the epoch of observation[10,19]. To reproduce the SED and the size of the Balmer break, it is necessary to consider star formation histories extending about 300 Myr earlier, corresponding to the onset of galaxy formation at redshifts of $z \approx 15$ (Figs. 2 and 3). Our fiducial model has a dominant old component whose SFR declined, thereby reproducing the Balmer break, and a much younger component that reproduces the strength of both the Lyα and the [O III] 88-μm emission. Although the detailed parameters of these two components cannot be uniquely determined with the current data, we demonstrate in Methods that a dominant earlier phase of star formation whose activity subsequently declined is necessary to reproduce the Balmer break. Indeed, this early component contributed to the bulk of the stellar mass of MACS1149-JD1 observed at $z = 9.1$ ($M_* \approx (10/\mu) \times 10^9 M_\odot$; Fig. 3). Its onset at $z = 15.4^{+2.7}_{-2.0}$ thus represents the logical epoch of formation of MACS1149-JD1[6].

During this dominant first episode of star formation, feedback processes probably weaken or even terminate star formation activity, leading to a quiescent phase until a second episode, when gaseous inflow can rejuvenate star formation at $z = 9.1$, leading to the detection of [O III] 88-μm emission. Intense ultraviolet radiation from this early generation of massive stars may create an ionized bubble[20–22] surrounding MACS1149-JD1, displacing the damping wing of Lyα absorption by intergalactic neutral hydrogen bluewards by ~500 km s^{-1}, thus facilitating transmission through the predominantly neutral intergalactic medium. Although this model is somewhat speculative, given its tentative nature, inflowing gas may then provide an explanation for the blueshift of Lyα with respect to the [O III] 88-μm line by several hundred kilometres per second[23,24] (see Methods).

MACS1149-JD1 is already a well established galaxy. Although we are observing a secondary episode of star formation at $z = 9.1$, the galaxy formed the bulk of its stars at a much earlier epoch. Our results indicate that it may be feasible to directly detect the earliest phases of galaxy formation, beyond the redshift range currently probed with the HST, with future facilities such as the James Webb Space Telescope. In addition, our observations demonstrate the great power of ALMA[19] to spectroscopically identify galaxies at $z > 9$, showing that it will also have a central role in investigating first-generation galaxies.

Online content

Any Methods, including any statements of data availability and Nature Research reporting summaries, along with any additional references and Source Data files, are available in the online version of the paper at https://doi.org/10.1038/s41586-018-0117-z.

Received: 24 November 2017; Accepted: 14 March 2018;
Published online 16 May 2018.

1. Robertson, B. E., Ellis, R. S., Furlanetto, S. R. & Dunlop, J. S. Cosmic reionization and early star-forming galaxies: a joint analysis of new constraints from Planck and the Hubble Space Telescope. *Astrophys. J.* **802**, L19 (2015).
2. Bouwens, R. et al. Reionization after Planck: the derived growth of the cosmic ionizing emissivity now matches the growth of the galaxy UV luminosity density. *Astrophys. J.* **811**, 140 (2015).
3. Stark, D. Galaxies in the first billion years after the Big Bang. *Annu. Rev. Astron. Astrophys.* **54**, 761–803 (2016).
4. McLeod, D., McLure, R. & Dunlop, J. The $z = 9-10$ galaxy population in the Hubble Frontier Fields and CLASH surveys: the $z = 9$ luminosity function and further evidence for a smooth decline in ultraviolet luminosity density at $z \geq 8$. *Mon. Not. R. Astron. Soc.* **459**, 3812–3824 (2016).
5. Oesch, P. et al. The dearth of $z \sim 10$ galaxies in all HST legacy fields – the rapid evolution of the galaxy population in the first 500 Myr. *Astrophys. J.* **855**, 105 (2018).
6. Zheng, W. et al. A magnified young galaxy from about 500 million years after the Big Bang. *Nature* **489**, 406–408 (2012).
7. Zheng, W. et al. Young galaxy candidates in the Hubble Frontier Fields. IV. MACS J1149.5+2223. *Astrophys. J.* **836**, 210 (2017).
8. Huang, K. et al. Spitzer UltRa Faint SUrvey Program (SURFS UP). II. IRAC-detected Lyman-break galaxies at $6<z<10$ behind strong-lensing clusters. *Astrophys. J.* **817**, 11 (2016).
9. Kawamata, R. et al. Precise strong lensing mass modeling of four Hubble Frontier Field clusters and a sample of magnified high-redshift galaxies. *Astrophys. J.* **819**, 114 (2016).
10. Inoue, A. K. et al. Detection of an oxygen emission line from a high-redshift galaxy in the reionization epoch. *Science* **352**, 1559–1562 (2016).
11. da Cunha, E. et al. On the effect of the cosmic microwave background in high-redshift (sub-)millimeter observations. *Astrophys. J.* **766**, 13 (2013).
12. Hildebrand, R. H. The determination of cloud masses and dust characteristics from submillimetre thermal emission. *Q. J. R. Astron. Soc.* **24**, 267–282 (1983).
13. Smit, R. et al. High-precision photometric redshifts from Spitzer/IRAC: extreme [3.6] – [4.5] colors identify galaxies in the redshift range $z \sim 6.6 - 6.9$. *Astrophys. J.* **801**, 122 (2015).
14. Roberts-Borsani, G. et al. $z \geq 7$ galaxies with red Spitzer/IRAC [3.6]–[4.5] colors in the full CANDELS data set: the brightest-known galaxies at $z \sim 7-9$ and a probable spectroscopic confirmation at $z = 7.48$. *Astrophys. J.* **823**, 143 (2016).
15. Stark, D. et al. Lyα and C III] emission in $z = 7-9$ galaxies: accelerated reionization around luminous star-forming systems? *Mon. Not. R. Astron. Soc.* **464**, 469–479 (2017).
16. Hoag, A. et al. HST grism observations of a gravitationally lensed redshift 9.5 galaxy. *Astrophys. J.* **854**, 39 (2018).
17. Carniani, S. et al. Extended ionised and clumpy gas in a normal galaxy at $z = 7.1$ revealed by ALMA. *Astron. Astrophys.* **605**, A42 (2017).
18. Vernet, J. et al. X-shooter, the new wide band intermediate resolution spectrograph at the ESO Very Large Telescope. *Astrophys. J.* **536**, A105 (2011).

19. Inoue, A. K. et al. ALMA will determine the spectroscopic redshift $z>8$ with FIR [O III] emission lines. *Astrophys. J.* **780**, L18 (2013).
20. Cen, R. & Haiman, Z. Quasar Strömgren spheres before cosmological reionization. *Astrophys. J.* **542**, L75–L78 (2000).
21. Haiman, Z. The detectability of high-redshift Lyα emission lines prior to the reionization of the Universe. *Astrophys. J.* **576**, L1–L4 (2002).
22. Hu, E. et al. An ultraluminous Lyα emitter with a blue wing at $z=6.6$. *Astrophys. J.* **825**, L7 (2016).
23. Dijkstra, M., Haiman, Z. & Spaans, M. Lyα radiation from collapsing protogalaxies. II. Observational evidence for gas infall. *Astrophys. J.* **649**, 14–36 (2006).
24. Verhamme, A. et al. Using Lyman-α to detect galaxies that leak Lyman continuum. *Astron. Astrophys.* **578**, A7 (2015).

Acknowledgements We thank K. Umetsu for a discussion of gravitational lensing models; K. Nakanishi, F. Egusa and K. Saigo for discussions about handling ALMA data; S. Kikuchihara for supporting MOSFIRE observations; and H. Yajima and A. Zitrin for discussions. We acknowledge support from: NAOJ ALMA Scientific Research Grant number 2016-01 A (T.H. and A.K.I.); European Research Council Advanced Grant FP7/669253 (N.L. and R.S.E.) and 339177 (C.E.R.); KAKENHI grants 26287034 and 17H01114 (K.M. and A.K.I.), 17H06130 (Y. Tamura), 17H04831 (Y.M.), 16H01085 (T.O.), 16H02166 (Y. Taniguchi), 15K17616 (B.H.), 17K14252 (H.U.), JP17H01111 (I.S.), 16J03329 (Y.H.) and 15H02064 (M.O.); the grant CONICYT-Chile Basal-CATA PFB-06/2007, FONDECYT Regular 1141218 (F.E.B.); NAOJ Visiting Fellow Program (N.H.H.). ALMA is a partnership of ESO (representing its member states), NSF (USA) and NINS (Japan), together with NRC (Canada), NSC and ASIAA (Taiwan) and KASI (Republic of Korea), in cooperation with the Republic of Chile. The Joint ALMA Observatory is operated by ESO, AUI/NRAO and NAOJ. This work incorporates observations made with ESO Telescopes at the La Silla Paranal Observatory. This work is also partly based on observations made with the Spitzer Space Telescope, which is operated by the Jet Propulsion Laboratory, California Institute of Technology under a contract with NASA, as well as observations obtained with the NASA/ESA Hubble Space Telescope at the Space Telescope Science Institute (STScI), which is operated by the Association of Universities for Research in Astronomy, Inc. under NASA contract NAS 5-26555.

Author contributions T.H., N.L., R.S.E., and A.K.I. wrote the paper. T.H. and Y. Tamura reduced and analysed ALMA data. T.H. produced Figs. 1 and 2 and Extended Data Figs. 1, 3–5. N.L. reduced and analysed X-shooter data and produced Extended Data Figs. 2 and 7. K.M. and E.Z. performed SED fitting analyses. K.M. produced Fig. 3 and Extended Data Fig. 6. W.Z. carried out the astrometry analysis on HST and IRAC data. N.L. and C.E.R. performed lensing analyses. H.M., I.S., T.O., N.Y., Y. Taniguchi, B.H., H.U. and Y.M. contributed to the ALMA observational strategy. N.H.H. independently inspected the ALMA data. G.R.B., T.F. and R.P. inspected independently the X-shooter spectra. G.R.B. contributed to the observations. F.E.B. contributed to the X-shooter observational strategy. Y.H. and M.O. performed MOSFIRE observations and analysed the archival and our own MOSFIRE data. All authors discussed the results and commented on the manuscript.

Competing interests The authors declare no competing interests.

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0117-z.

Reprints and permissions information is available at http://www.nature.com/reprints.

Correspondence and requests for materials should be addressed to T.H.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

Cosmological model. The adopted cosmological parameters[25] are $H_0 = 70.4\,\mathrm{km\,s^{-1}\,Mpc^{-1}}$ (Hubble parameter), $\Omega_\mathrm{m} = 0.272$ (total matter density), $\Omega_\mathrm{b} = 0.045$ (baryon density) and $\Omega_\Lambda = 0.728$ (cosmological constant).

ALMA observations and data reduction. We observed MACS1149-JD1 with ALMA in Band 7 with a configuration C40-3 (project 2015.1.00428.S; principal investigator, A. K. Inoue). To cover the uncertainty derived from the photometric redshift analysis[4,6–9], we used four setups with contiguous frequencies, labelled as T3, T4, T5 and T6, encompassing the frequency range 314.4–340.5 GHz and the redshift range $z = 9.0$–9.8. In each setup, we used a total bandwidth of 7.5 GHz split into four spectral windows, each with a bandwidth of 1.875 GHz in the frequency division mode. Each spectral window has a resolution of 7.8125 MHz, corresponding to a velocity resolution of $\sim 7\,\mathrm{km\,s^{-1}}$. The total observation times are 75.6 min, 35.3 min, 119.4 min and 42.3 min for T3, T4, T5 and T6, respectively. The T3, T4, T5 and T6 data were reduced using the CASA pipeline versions 4.7.0, 4.5.2, 4.7.2 and 4.6.0, respectively, with a standard calibration script provided by the ALMA observatory. We then produced final images and cubes with the CLEAN task using natural weighting to maximize point-source sensitivity. The spatial resolution was $0.62'' \times 0.52''$ (FWHM) and the beam position angle was −8.9°. A quasar, J1229 + 0203, was used for bandpass and flux calibrations, for which the flux uncertainty was estimated to be smaller than ~10%.

[O III] 88-μm line. To search for a line, we created a data cube, six native channels of which were binned, resulting in a velocity resolution of $\sim 42\,\mathrm{km\,s^{-1}}$. In the T5 setup at ~335 GHz, we found a 3.0σ signal in five continuous binned channels, where 1σ is the local noise estimated with the CASA task *imstat*. This frequency region is free from atmospheric absorption features. We then created a velocity-integrated intensity image between 335.5 GHz and 335.8 GHz. The peak intensity of MACS1149-JD1 is $129.8 \pm 17.5\,\mathrm{mJy\,km\,s^{-1}}$ per beam, corresponding to a significance level of 7.4σ.

The spatial centroid of the emission line is in good positional agreement with that of the ultraviolet continuum emission observed by HST (Fig. 1). Both images are similarly elongated along the gravitational lensing shear. We measured the integrated line flux using the CASA task *imfit* to be $0.229 \pm 0.048\,\mathrm{Jy\,km\,s^{-1}}$. To obtain the redshift, we extracted the one-dimensional spectrum from the region with 3σ signals in the velocity-integrated intensity image. As can be seen, the [O III] line is detected at around 335.6 GHz (or 893.2 μm) in the Solar System barycentric frame. By applying a Gaussian fit to the line and with a rest-frame [O III] frequency of 3393.006244 GHz, we obtain a redshift $z_{[\mathrm{OIII}]} = 9.1096 \pm 0.0006$ and an FWHM of $154 \pm 39\,\mathrm{km\,s^{-1}}$, which is reasonable for a low-mass galaxy[26]. The integrated flux and redshift leads to an observed luminosity of $(7.4 \pm 1.6)(10/\mu) \times 10^7\,L_\odot$.

Using the CASA task *imfit*, we obtained the deconvolved size of MACS1149-JD1 as $(0.82'' \pm 0.25'') \times (0.30'' \pm 0.14'')$. Assuming that lensing effects are equal for the major and minor axes, the intrinsic size is $(3.7 \pm 1.1)\mu^{-0.5}\,\mathrm{kpc} \times (1.4 \pm 0.9)\mu^{-0.5}\,\mathrm{kpc}$.

Upper limits on dust continuum emission, total infrared luminosity and dust mass. To create a dust continuum image, we collapsed (that is, summed up along the frequency axis) all four setup data in the frequency range from 314.4 GHz to 340.5 GHz. In this procedure, we excluded a frequency range around the [O III] line with a width of $\pm 3\mathrm{FWHM}_{[\mathrm{OIII}]}$. The CASA task *imstat* gave the r.m.s. noise level for the continuum image as 17.7 μJy per beam. Extended Data Fig. 1a shows ALMA dust continuum contours of MACS1149-JD1 overlaid on the HST F160W image. In the continuum image, we do not detect any signal above 3σ, placing a stringent upper limit on dust continuum emission at $S_{\nu,90\mu\mathrm{m}} < 5.3(10/\mu)\,\mu\mathrm{Jy}$ per beam.

We estimate the total infrared luminosity by integrating the modified blackbody radiation over 8–1,000 μm with an emissivity index of $\beta_\mathrm{d} = 1.5$ and a dust temperature of $T_\mathrm{d} = 30$–$60\,\mathrm{K}$, similarly to previous studies[27–30]. The results, corrected for cosmic microwave background effects[11,31], are listed in Extended Data Table 1. Assuming a dust mass absorption coefficient of $\kappa = \kappa_0(\nu/\nu_0)^{\beta_\mathrm{d}}$, where $\kappa_0 = 10\,\mathrm{cm^2\,g^{-1}}$ at 250 μm (ref. 12), we obtain a 3σ dust mass upper limit of $5.3(10/\mu) \times 10^5\,M_\odot$ for the case of 40 K. The upper limit on the dust mass is increased if the dust temperature is lowered (see Extended Data Table 1). In this estimation, we have assumed the size of the dust-emitting region to be a single beam size, following other high-*z* null-detection cases[10,31–33]. If, instead, we assume the same size as that of the [O III]-emitting region, which is roughly three times larger than the beam size, the upper limits can be relaxed by a factor of $\sqrt{3}$ compared with the values listed in Extended Data Table 1.

VLT and X-shooter observations and results. We performed the X-shooter observations in February and April 2017 (programme 098.A-0534; principal investigator, N. Laporte) for 4.25 h and 4 h, respectively. We adopted a nodding offset of 4.0″ so that a bright galaxy entered the slit in every other exposure. By detecting the continuum in these half-exposures, we confirmed that MACS1149-JD1 was correctly acquired. The chosen slit orientation (+15°) encompasses both the bright galaxy and the elongated shape of the lensed target (Extended Data Fig. 2). To maximise efficiency, we used 900-s, 810-s and 740-s exposures in the NIR, VIS and UVB arms of X-shooter, respectively.

The data were reduced using the European Southern Observatory (ESO) Reflex tool[34] (version 2.8) with the latest version of the X-shooter pipeline (version 2.9.3). The final spectrum was obtained by combining all reduced exposures using both EsoReflex and a master flat derived from all the data, as well as the IRAF *imcombine* task on the individual observing blocks. Both methods give similar results. No emission line was found in the UVB- and VIS-arm spectra. Visual inspection of the NIR-arm spectrum shows a possible emission line at 12,267.4 Å with an integrated flux of $(4.3 \pm 1.1) \times 10^{-18}\,\mathrm{erg\,s^{-1}\,cm^{-2}}$. On the basis of the r.m.s. value measured in similarly sized adjacent apertures, the emission line has a formal significance of 4σ. Taking into account air refraction and the motion of the observatory, we deduce a vacuum wavelength in the barycentric system of 12,271.51 ± 2.27 Å, corresponding to Lyα at a redshift of 9.0942 ± 0.0019. We rule out other identifications (such as [O II], Hβ, [O III] and Hα) because further emission lines would be seen, given our wide wavelength coverage and typical line ratios seen in other star-forming sources. Within the likely wavelength range of Lyα, the ratio of resolution elements with positive and negative deviations above a signal-to-noise ratio of 4 exceeds 5, emphasizing that the detection cannot be dismissed as a fluctuation in the noise variation. Although some Keck MOSFIRE data exist for MACS1149-JD1, both in the Keck archive and from a recent campaign, these data have insufficient additional exposure time to strengthen the signal-to-noise ratio of the X-shooter detection.

The redshifts obtained from [O III] and Lyα are slightly different, corresponding to a velocity offset of $\Delta v_{\mathrm{Ly}\alpha} = -450 \pm 60\,\mathrm{km\,s^{-1}}$ (Extended Data Fig. 3). For a single galaxy emitting both [O III] and Lyα, inflowing gas with a high neutral hydrogen column density, $\log[N_\mathrm{HI}\,(\mathrm{cm^{-2}})] \approx 20$–$21$ would lead to Lyα becoming blueshifted[23,24], and similar offsets have been seen[35] at $z \approx 2$–3. Such inflowing gas would rejuvenate MACS1149-JD1, causing a secondary burst after a relatively inactive phase, as indicated by our SED analyses. The Lyα line profile would also be affected by the surrounding intergalactic medium. Because our SED analysis indicates a dominant phase of earlier star formation, we may expect a large ionized bubble around MACS1149-JD1. Assuming a simple model[20–22], the Strömgren radius can be estimated as $R_\mathrm{s} = (3N_\mathrm{ion}/4\pi\langle n_\mathrm{H}\rangle)^{1/3}$, where N_ion is the total number of ionizing photons emitted by the early stellar component and $\langle n_\mathrm{H}\rangle$ is the mean hydrogen density at $z = 9.1$. From the SED analyses we estimate $N_\mathrm{ion} \approx 10^{70} \times (10/\mu)$, regardless of the star formation history (SFH). Assuming an escape fraction of 20%, we obtain a Strömgren radius of $0.4(f_\mathrm{esc}/0.2)^{1/3}(10/\mu)^{1/3}$ Mpc, corresponding to a velocity offset of $\sim 500(f_\mathrm{esc}/0.2)^{1/3}(10/\mu)^{1/3}\,\mathrm{km\,s^{-1}}$, where f_esc corresponds to the escape fraction of ionizing photons from galaxies to the intergalactic medium.

Alternatively, it is possible that MACS1149-JD1 comprises two $z = 9.1$ galaxies separated kinematically by $450\,\mathrm{km\,s^{-1}}$. In this case, the IRAC excess could be associated with either galaxy; regardless, the Lyα-emitting component would lie within the ionized bubble. Indeed, we observe a $z = 7.1$ galaxy that has the Lyα, [O III] and [C II] lines, each offset spatially and in velocity space[17].The meaning of the sentence is different from our original intention. The results mentioned in this sentence is not results from us or our team members but from Carniani et al. (2018; ref 17).Therefore, please replace the sentence with the new one:Indeed, a recently observed z=7.1 galaxy shows Lyα, [O III] and [C II] lines which are each offset spatially and in velocity space[17]. This galaxy shows a Lyα line blueshifted by $\sim 500\,\mathrm{km\,s^{-1}}$ with respect to [O III].

Astrometry. Several studies have demonstrated that there are, in some cases, spatial offsets between ALMA-detected sources and their HST counterparts[29,36]. We took advantage of a dusty spiral galaxy, CLASH 2882[37–39], in our ALMA field of view to examine possible astrometric offsets. Extended Data Fig. 1b shows ALMA dust continuum contours of CLASH 2882 overlaid on the HST F160W image. Dust continuum is clearly detected at the exact position of the near-infrared counterpart of CLASH 2882 at a significance level of $\sim 10\sigma$ (where $1\sigma = 16.0$ μJy per beam). The centroids of ALMA and HST are consistent within 0.1″.

As can be seen from the HST morphology shown in Fig. 1 and Extended Data Fig. 1a, MACS1149-JD1 is elongated and may be formed by two components at $z = 9.1$ separated by 0.33″. To determine the exact position of the Lyα-emitting region, we collapsed the final two-dimensional spectrum in the wavelength direction and measured the separation between Lyα and the continuum of the reference galaxy. Within a positional uncertainty of ~0.7″, the Lyα-emitting region could arise in either component. As revealed in Figs. 2 and 3, the strong IRAC excess at 4.5 μm is a prominent feature of MACS1149-JD1. To determine from which component the IRAC 4.5-μm emission arises, we first examined the astrometry differences between the HST and IRAC 4.5-μm emissions using bright point-source stars. Within a 1σ uncertainty of ~0.3″ and the resolution of the IRAC camera, the centroid of the 4.5-μm channel is consistent with emission from either of the two HST counterparts.

Magnification factor. With our new spectroscopic redshift, we revisited the magnification calculations using all the mass models in the framework of the Frontier Fields survey[40–46]. The magnification values range from $\mu = 4.9^{+0.4}_{-0.5}$ (Diego version 4.1) to $\mu = 89.0^{+192.7}_{-40.8}$ (CATS version 4.1) with a mean value of $\langle\mu\rangle = 21.8^{+260.0}_{-17.4}$.

However, if we remove the most extreme values, the mean magnification factor decreases to $\langle\mu\rangle = 11.7^{+80.1}_{-5.1}$, where in the errors we have taken into account the dispersion of the values obtained from all the Frontier Fields models. For simplicity, in this study we assume a fiducial value of 10 for the magnification factor.

SED modelling. We built the SED of our target by combining data from the three deep HST images obtained within the framework of the Frontier Fields survey[7], the deep K_s image obtained with VLT/HAWK-I[16], as well as the deep Spitzer/IRAC 3.6-μm and 4.5-μm data[7]. In addition, we used our deep upper limit on dust continuum emission, as well as our measurement of the [O III] flux described above.

We used our own SED fitting code[47]. Briefly, this code uses the stellar population synthesis model of GALAXEV[48], including nebular continuum and emission lines[49]. The [O III] line flux was estimated from the metallicity and SFR using semi-empirical models[10,19]. The ionizing photon escape fraction was fixed to zero. We assumed Calzetti's law[50] for dust attenuation and adopted empirical dust-emission templates[51]. We used a Chabrier initial-mass function[52] at $0.1\,M_\odot - 100\,M_\odot$ and applied a mean intergalactic medium model[53]. The model parameters and their steps are summarized in Extended Data Table 2. To estimate the best-fit parameters, we used the least-χ^2 formula[54] including an analytic treatment of upper limits for non-detection. Uncertainties on the parameters were estimated on the basis of a Monte Carlo technique.

In a first attempt, we adopted a single exponentially declining SFH. Extended Data Fig. 4a shows the best-fit SED obtained with a reduced χ^2 of $\chi^2_\nu \approx 6.9$ and a stellar age of ~300 Myr. Although this stellar age is comparable to that obtained by previous studies[7,16], this model could not reproduce the [O III] line flux observed with ALMA. The situation is worse if we assume a constant SFH with a reduced χ^2 of $\chi^2_\nu \approx 11.4$, which leads to a stellar age of ~500 Myr (Extended Data Fig. 4b).

We next considered a two-stellar-component SED comprising a young starburst and an old population. To reproduce the observed SED with as small a number of parameters as possible, we assumed constant SFRs for both components. Extended Data Fig. 5 shows a schematic overview of our models. If one assumes an exponentially declining (rising) SFH for the old component, there is always an equivalent constant-SFH model with a lower (higher) age of the population. We adopted several star formation durations for the old component, namely, $\tau = 10$ Myr, 100 Myr, 200 Myr, 300 Myr and 400 Myr. In these two-component models, we can reproduce the observations with $\chi^2_\nu \approx 2.8-3.0$. Models with $\tau = 300$ Myr and 400 Myr lead to a stellar age of 400 Myr at $z = 9.1$, similar to the age of the Universe at this redshift, which we consider physically unacceptable. The other three models give a similar-quality fit and comparable ages for the old stellar population, ranging from 230 Myr to 320 Myr. Owing to the similarity of the physical and statistical results produced by the three remaining models, we cannot easily discriminate between them. For this reason, we select the $\tau = 100$ Myr model as our fiducial case but include uncertainties that reflect the standard deviation caused by the choice of the model with $\tau = 10$ Myr or $\tau = 200$ Myr (Extended Data Table 3). Extended Data Fig. 4c and d shows the two best-fit SEDs for $\tau = 10$ Myr and 200 Myr.

Although we can exclude contamination of IRAC channel 2 from the [O III] 5,007-Å line at the [O III] redshift $z = 9.1096$ (and the lower Lyα redshift), there may be contamination[13] from intense Hβ and [O III] 4,959-Å emission. Our nebular emission line model[49] already accounts for such contributions up to a rest-frame equivalent width (Hβ plus [O III] 4,959 Å) of ~1,550 Å in young metal-poor cases. Nevertheless, Extended Data Fig. 6 shows that such young metal-poor models cannot reproduce the red IRAC colour because of the contribution of strong nebular continuum and [O II] 3,727-Å line emission in IRAC channel 1. This explanation for the IRAC excess is further disfavoured by the observed dust emission upper limit (Extended Data Fig. 6c).

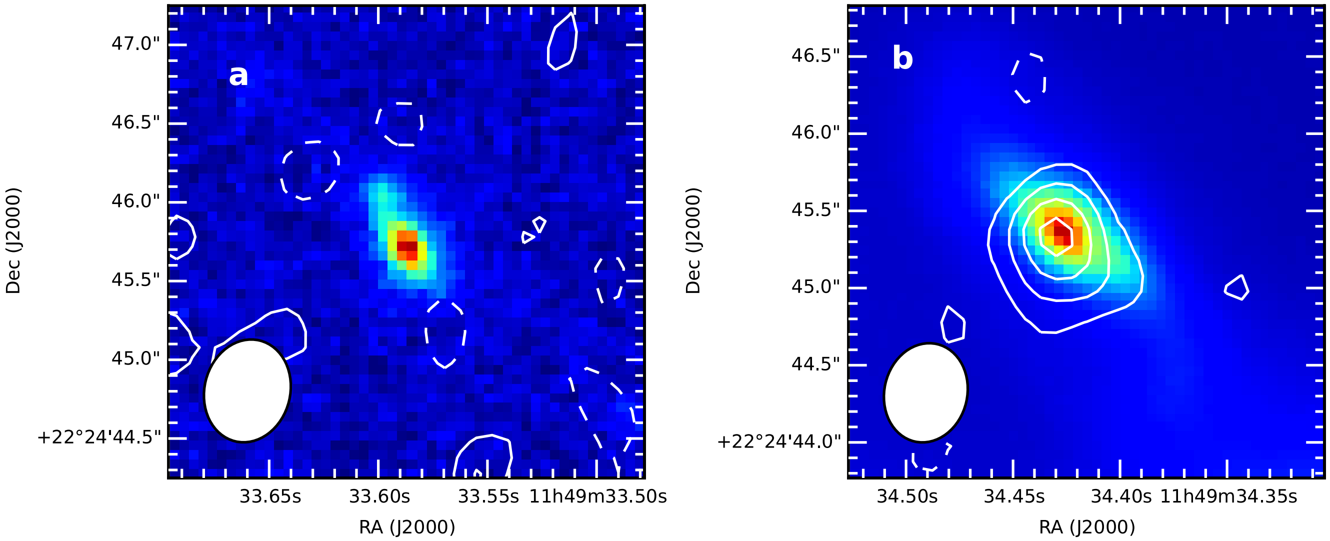
Dynamical mass. We derive the dynamical mass, $M_{\rm dyn}$, of MACS1149-JD1 following analyses[55] at $z \approx 2$. Because MACS1149-JD1 does not show a velocity gradient in [O III], we assume that MACS1149-JD1 is dispersion-dominated. In this case, the dynamical mass can be obtained as $M_{\rm dyn} = 6.7(\sigma_{\rm line})^2 r_{1/2}/G$, where $\sigma_{\rm line}$ is the line velocity dispersion, $r_{1/2}$ is the half-light radius and G is the gravitational constant. Taking the lensing effect into account, the intrinsic half-light radius becomes $(1.9 \pm 0.6)\mu^{-0.5}$ with an uncertainty of $\mu^{-0.5}$ depending on the direction of elongation. We thus obtain $M_{\rm dyn} = (4 \pm 3)(10/\mu)^{0.5} \times 10^9\,M_\odot$. Therefore, the dynamical mass is comparable to or larger than the stellar mass, $(1.1 \pm 0.5)(10/\mu) \times 10^9\,M_\odot$.

Predictions for future observations with the James Webb Space Telescope. The SFH deduced from the SED fits allows us to trace the earlier luminosity evolution of MACS1149-JD1 as a function of cosmic age. Accordingly, we computed the SED from $z = 15$ to $z = 9.1$ and found that the source would have been as bright as $m_{1500} = 26$ AB at $z = 12$ (corrected for magnification; Extended Data Fig. 7). Such a source would be easily detectable by the James Webb Space Telescope. Between $z \approx 12$ and $z \leq 15$, it would be detected with the near-infrared camera (NIRCam) at 10σ within 20 min. Furthermore, according to the sensitivity of the near-infrared spectrograph (NIRSpec), the ultraviolet continuum and a significant Lyα break of the progenitor of a source such as MACS1149-JD1 would be detectable at a 5σ level in less than an hour (Extended Data Fig. 7).

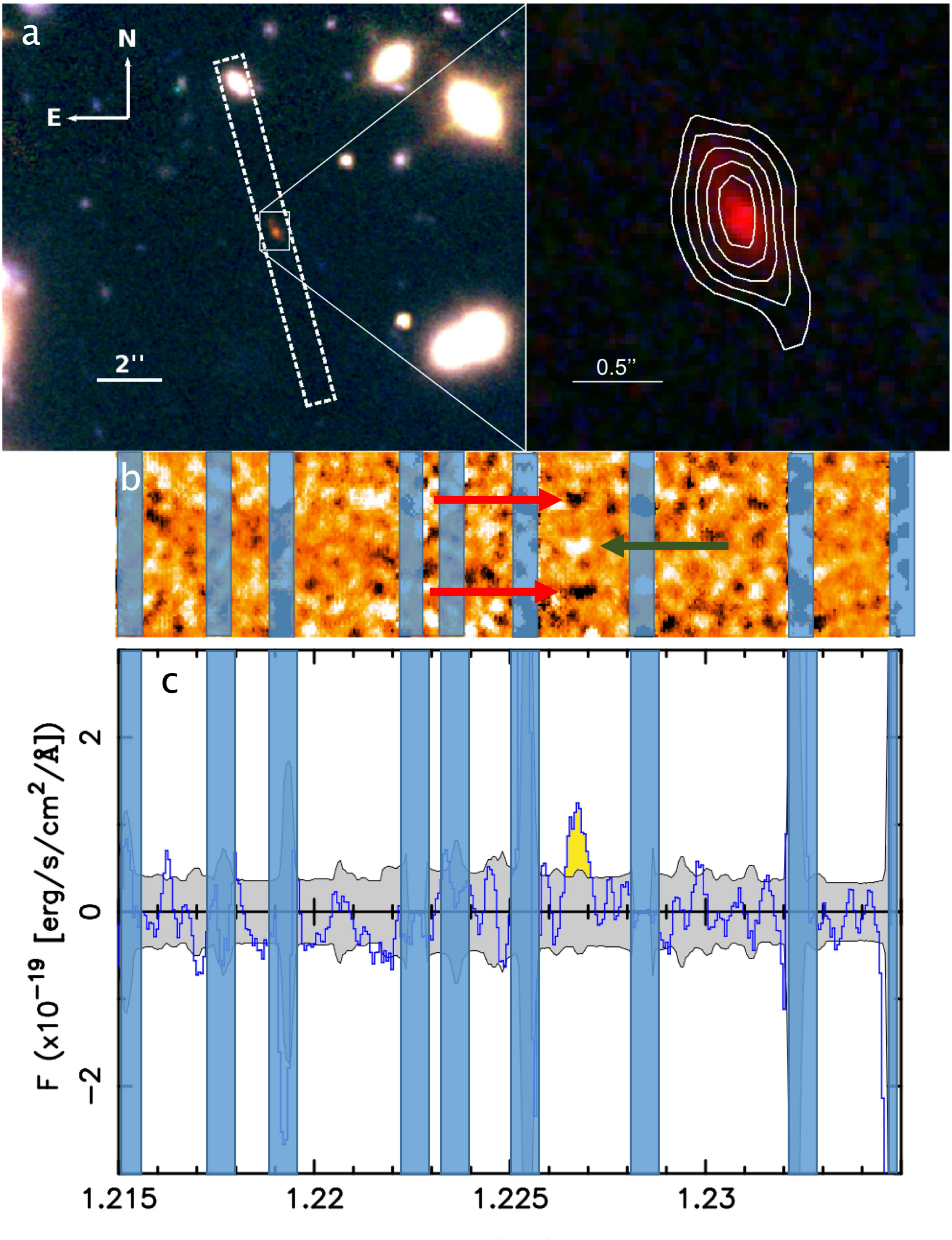
Code availability. The ALMA data were reduced using the CASA pipeline versions 4.7.0, 4.5.2, 4.7.2 and 4.6.0, for the T3, T4, T5 and T6 data, respectively. For SED fitting, we used a custom-made code, which is available at https://www.astr.tohoku.ac.jp/~mawatari/KENSFIT/KENSFIT.html with instructions.

Data availability. This work used the following ALMA data ADS/JAO.ALMA#2015.1.00428.S, which are available at https://almascience.nao.ac.jp/aq/?project_code=2015.1.00428.S. This study was also based on observations made with the ESO Telescopes at the La Silla Paranal Observatory under programme 098.A-0534, which are available at http://archive.eso.org/wdb/wdb/eso/sched_rep_arc/query?progid=098.A-0534(A). Data from HST and Spitzer, as well as VLT/HAWK-I photometry data that support our findings, are available in the IOPscience repository under DOI 10.3847/1538-4357/aa5d55[7] and DOI 10.3847/1538-4357/aaa9c2[16], respectively. The datasets generated or analysed during this study are available from the corresponding author on reasonable request.

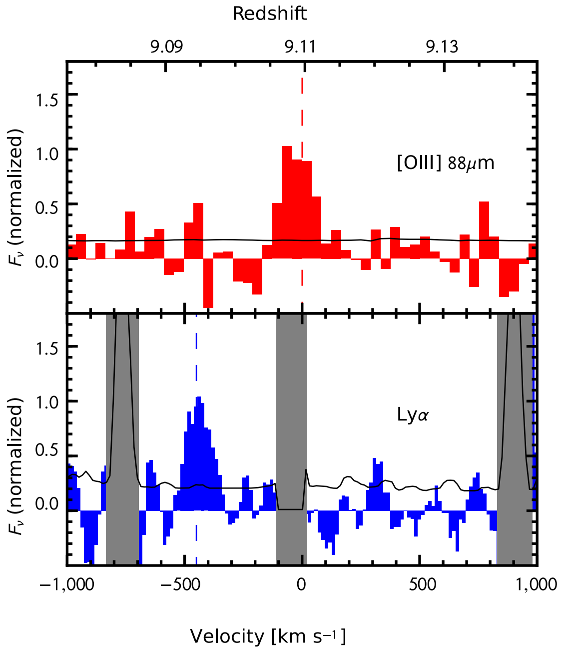
25. Komatsu, E. et al. Seven-year Wilkinson microwave anisotropy probe (WMAP) observations: cosmological interpretation. *Astrophys. J.* **192**, 18 (2011).
26. Hashimoto, T. et al. A close comparison between observed and modeled Lyα lines for $z \sim 2.2$ Lyα emitters. *Astrophys. J.* **812**, 157 (2015).
27. Watson, D. et al. A dusty, normal galaxy in the epoch of reionization. *Nature* **519**, 327–330 (2015).
28. Knudsen, K. K. et al. A merger in the dusty, $z = 7.5$ galaxy A1689-zD1? *Mon. Not. R. Astron. Soc.* **466**, 138–146 (2017).
29. Laporte, N. et al. Dust in the reionization era: ALMA observations of a $z = 8.38$ gravitationally lensed galaxy. *Astrophys. J.* **837**, L21 (2017).
30. Faisst, A. K. et al. Are high-redshift galaxies hot? Temperature of $z > 5$ galaxies and implications for their dust properties. *Astrophys. J.* **847**, 21 (2017).
31. Ota, K. et al. ALMA observation of 158 μm [C II] line and dust continuum of a $z = 7$ normally star-forming galaxy in the epoch of reionization. *Astrophys. J.* **792**, 34 (2014).
32. Ouchi, M. et al. An intensely star-forming galaxy at $z \sim 7$ with low dust and metal content revealed by deep ALMA and HST observations. *Astrophys. J.* **778**, 102 (2013).
33. Schaerer, D. et al. New constraints on dust emission and UV attenuation of $z = 6.5-7.5$ galaxies from millimeter observations. *Astron. Astrophys.* **574**, A19 (2015).
34. Freudling, W. et al. Automated data reduction workflows for astronomy. The ESO Reflex environment. *Astrophys. J.* **559**, A96 (2013).
35. Trainor, R., Steidel, C., Strom, A. & Rudie, G. The spectroscopic properties of Lyα-emitters at $z \sim 2.7$: escaping gas and photons from faint galaxies. *Astrophys. J.* **809**, 89 (2015).
36. González-López, J. et al. The ALMA Frontier Fields survey. I. 1.1 mm continuum detections in Abell 2744, MACS J0416.1−2403 and MACS J1149.5+2223. *Astron. Astrophys.* **597**, A41 (2017).
37. Dwek, E. et al. Dust formation, evolution, and obscuration effects in the very high-redshift universe. *Astrophys. J.* **788**, L30 (2014).
38. Dwek, E. et al. Submillimeter observations of CLASH 2882 and the evolution of dust in this galaxy. *Astrophys. J.* **813**, 119 (2015).
39. Zavala, J. A. et al. Early science with the Large Millimeter Telescope: dust constraints in a $z \sim 9.6$ galaxy. *Mon. Not. R. Astron. Soc.* **453**, L88–L92 (2015).
40. Richard, J. et al. Mass and magnification maps for the Hubble Space Telescope Frontier Fields clusters: implications for high-redshift studies. *Mon. Not. R. Astron. Soc.* **444**, 268–289 (2014).
41. Johnson, T. L. et al. Lens models and magnification maps of the six Hubble Frontier Fields clusters. *Astrophys. J.* **797**, 48 (2014).
42. Ishigaki, M. et al. Hubble Frontier Fields first complete cluster data: faint galaxies at $z \sim 5-10$ for UV luminosity functions and cosmic reionization. *Astrophys. J.* **799**, 12 (2015).
43. Keeton, C. R. On modeling galaxy-scale strong lens systems. *Gen. Relativ. Gravit.* **42**, 2151–2176 (2010).
44. Liesenborgs, J., De Rijcke, S. & Dejonghe, H. A genetic algorithm for the non-parametric inversion of strong lensing systems. *Mon. Not. R. Astron. Soc.* **367**, 1209–1216 (2006).
45. Diego, J. M., Protopapas, P., Sandvik, H. B. & Tegmark, M. Non-parametric inversion of strong lensing systems. *Mon. Not. R. Astron. Soc.* **360**, 477–491 (2005).
46. Merten, J. et al. Creation of cosmic structure in the complex galaxy cluster merger Abell 2744. *Mon. Not. R. Astron. Soc.* **417**, 333–347 (2011).
47. Mawatari, K. et al. Possible identification of massive and evolved galaxies at $z > 5$. *Publ. Astron. Soc. Jpn.* **68**, 46 (2016).
48. Bruzual, G. & Charlot, S. Stellar population synthesis at the resolution of 2003. *Mon. Not. R. Astron. Soc.* **344**, 1000–1028 (2003).
49. Inoue, A. K. Rest-frame ultraviolet-to-optical spectral characteristics of extremely metal-poor and metal-free galaxies. *Mon. Not. R. Astron. Soc.* **415**, 2920–2931 (2011).
50. Calzetti, D. et al. The dust content and opacity of actively star-forming galaxies. *Astrophys. J.* **533**, 682–695 (2000).
51. Rieke, H. et al. Determining star formation rates for infrared galaxies. *Astrophys. J.* **692**, 556–573 (2009).
52. Chabrier, G. Galactic stellar and substellar initial mass function. *Publ. Astron. Soc. Pacif.* **115**, 763–795 (2003).
53. Inoue, A. K. et al. An updated analytic model for attenuation by the intergalactic medium. *Mon. Not. R. Astron. Soc.* **442**, 1805–1820 (2014).
54. Sawicki, M. SEDfit: software for spectral energy distribution fitting of photometric data. *Publ. Astron. Soc. Pacif.* **124**, 1208–1218 (2012).
55. Förster Schreiber, M. et al. The SINS survey: SINFONI integral field spectroscopy of $z \sim 2$ star-forming galaxies. *Astrophys. J.* **706**, 1364–1428 (2009).



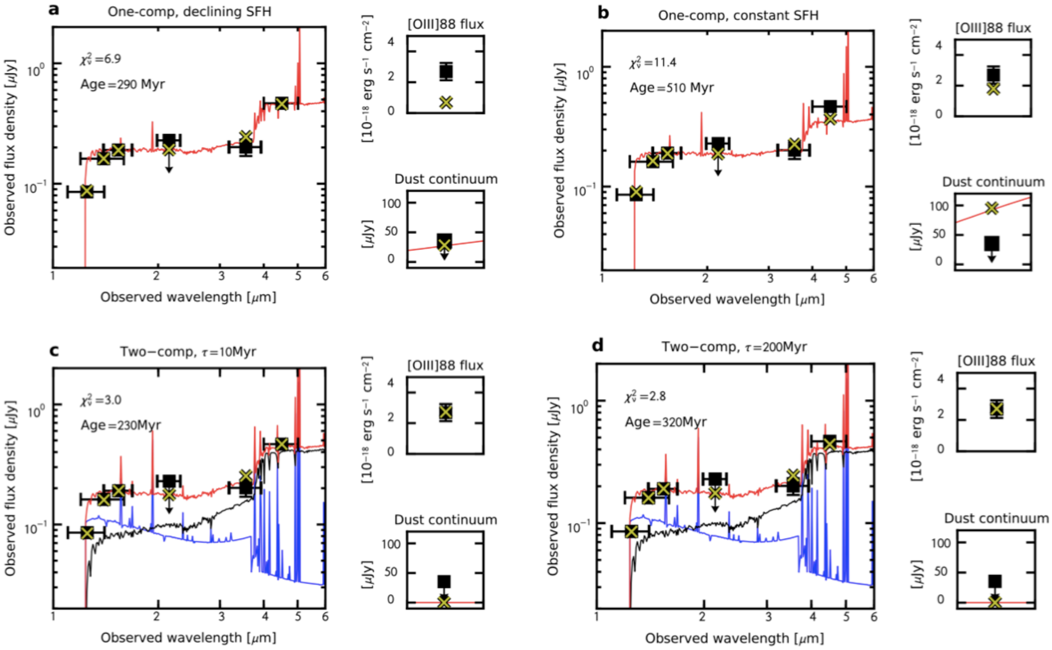
Extended Data Fig. 1 | ALMA dust contours of MACS1149-JD1 and a serendipitous continuum object. a, ALMA dust contours of MACS1149-JD1 overlaid on the HST F160W image. Contours are drawn at ± 2σ, where $\sigma = 17.7$ μJy per beam. Negative contours are shown by the dashed line. The ellipse at the lower left corner indicates the synthesized beam size of ALMA. **b**, Dust continuum of a dusty galaxy at $z = 0.99$ in our ALMA field of view, overlaid on the HST F160W image. Contours are drawn at -2σ, 2σ, 4σ, 6σ, 8σ and 9.5σ, where $\sigma = 16.0$ μJy per beam.



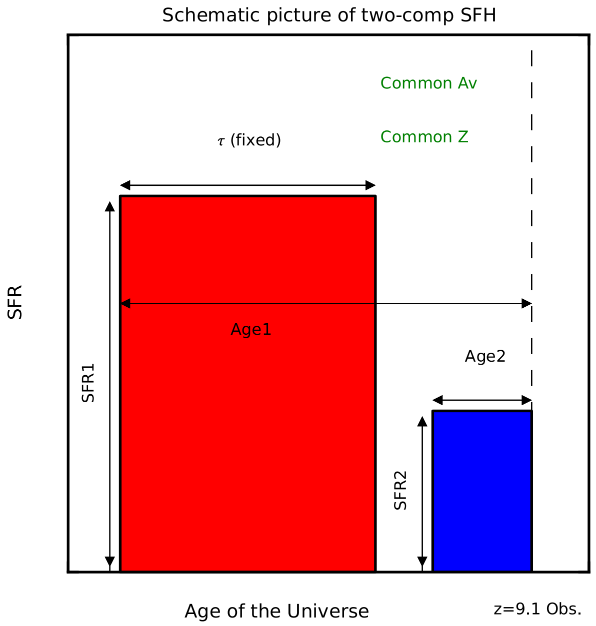
Extended Data Fig. 2 | X-shooter observations and Lyα spectra of MACS1149-JD1. a, Orientation of the X-shooter slit (white dashed-line rectangle), demonstrating the successful acquisition of MACS1149-JD1 data by alignment of the slit to follow the lensed elongation, as well as the inclusion of a bright foreground galaxy. **b**, Two-dimensional X-shooter spectra of MACS1149-JD1, with the position of Lyα marked with a green arrow and the two negative counterparts shown by red arrows. Sky lines are highlighted by blue rectangles. **c**, Extracted one-dimensional spectra in a 0.8″ aperture. Lyα is shaded in yellow, 2σ is shown in grey, and the sky lines are marked by blue rectangles.



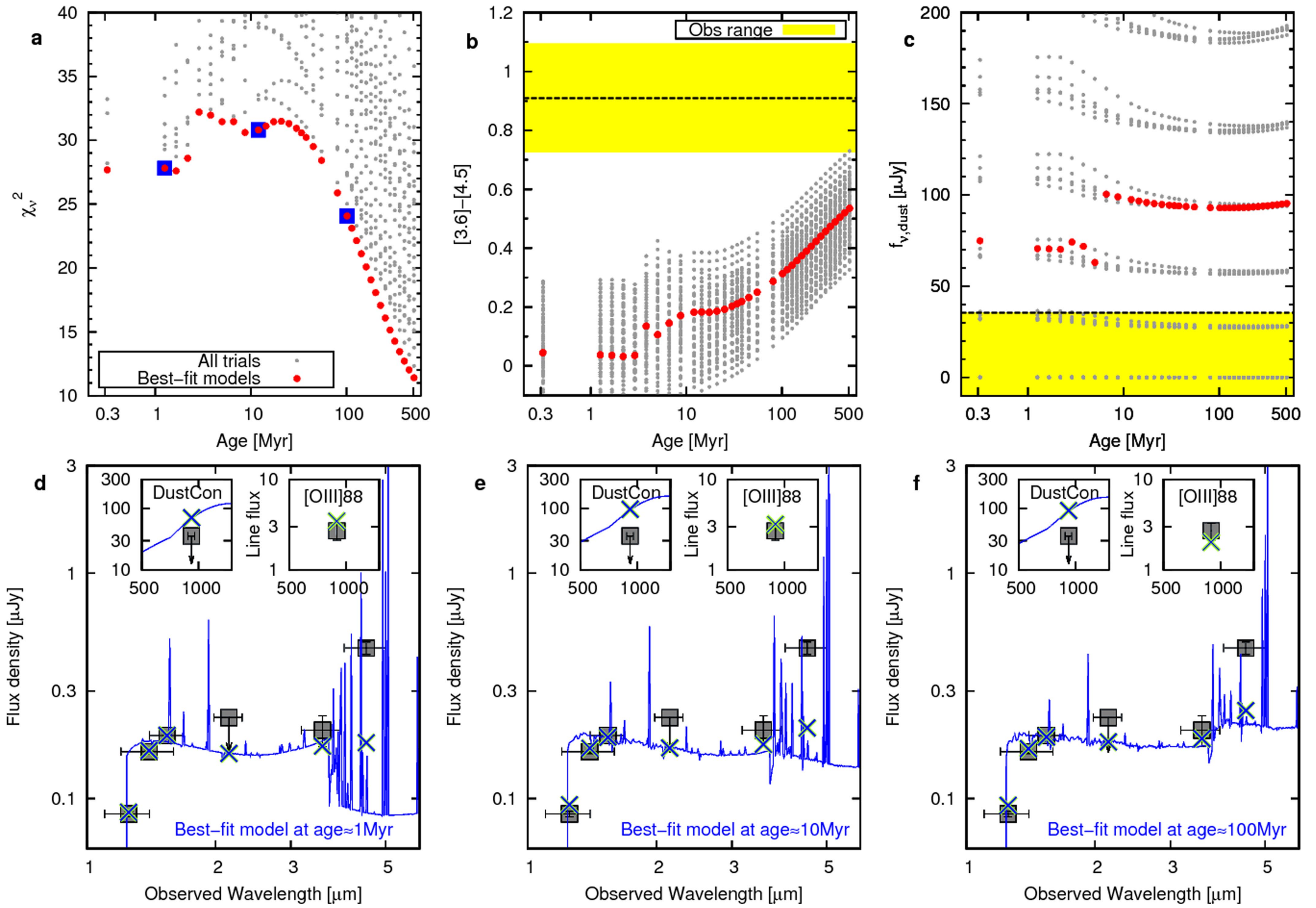
Extended Data Fig. 3 | ALMA [O III] 88-μm emission and X-shooter Lyα spectra in velocity space. The flux densities F_ν of [O III] and Lyα emissions are shown with a resolution of ~42 km s^{-1} and 15 km s^{-1}, respectively. The values are normalized by the peak flux densities. The zero-velocity point corresponds to the [O III] redshift, $z = 9.1096$ (red dashed line), and the Lyα offset is ~450 km s^{-1} (blue dashed line). Grey rectangles show regions contaminated by night-sky emission. The data at around −100 km s^{-1} to 0 km s^{-1} were removed from the analysis because the night-sky emission was too strong. The black solid lines indicate the 1σ noise level for these velocity resolutions.



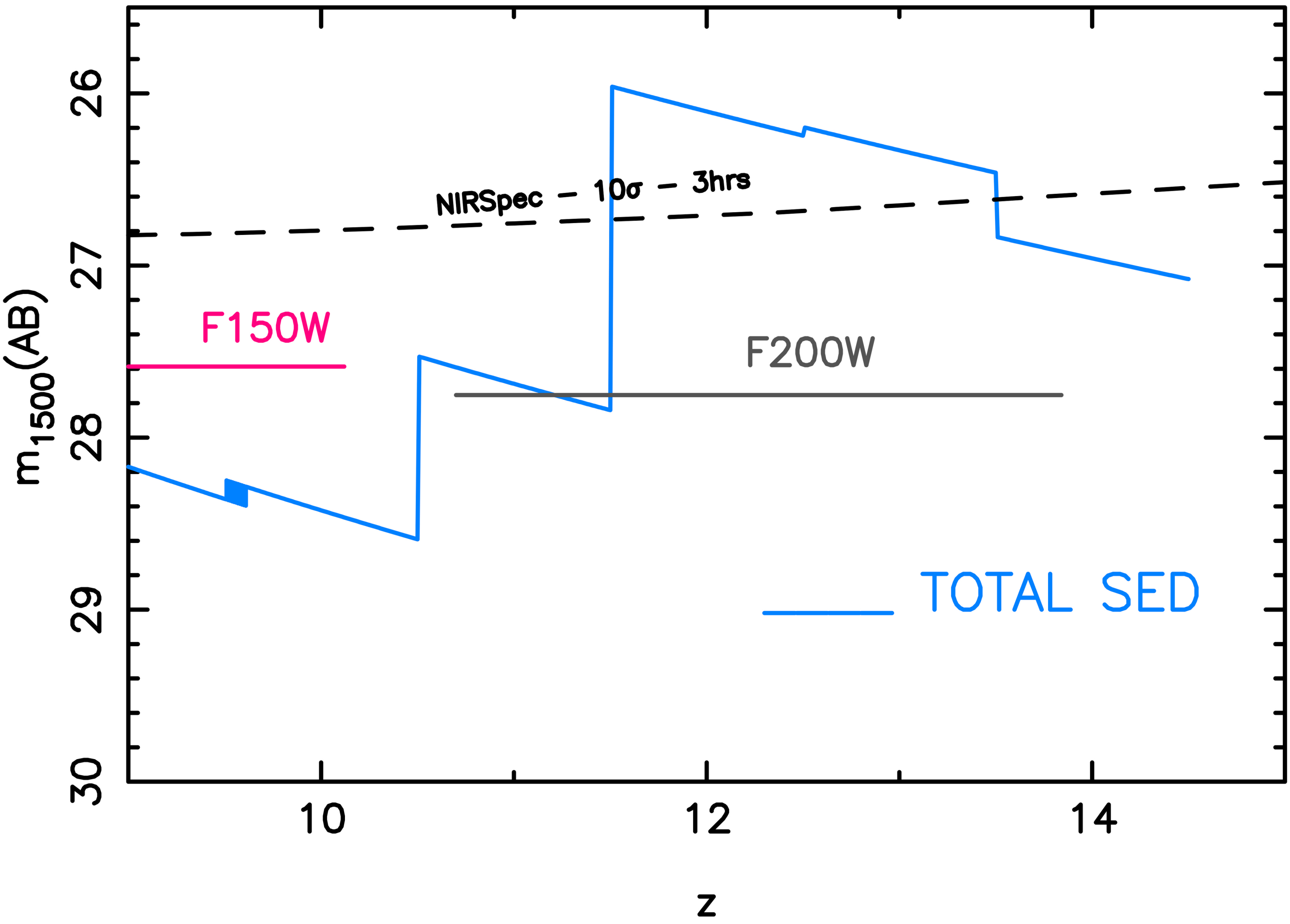
Extended Data Fig. 4 | Best-fit SEDs of MACS1149-JD1 with various SFHs. a, **b**, Best-fit SEDs obtained with a single stellar component, assuming an exponentially declining (**a**) and a constant (**b**) SFH. **c**, **d**, Best-fit SEDs obtained with two stellar components, assuming a constant SFH. The star formation duration of the old component is $\tau = 10$ Myr (**c**) and 200 Myr (**d**). The reduced χ^2 value, χ^2_ν, and the best-fit stellar age for each model is shown in the upper left corner. The meanings of the symbols are the same as those in Fig. 2.



Extended Data Fig. 5 | Schematic overview of the SFHs of our two-components models. The red and blue rectangles show the old and young stellar components with constant SFRs, respectively. The old component stops its star formation activity after a fixed duration, τ. The black vertical dashed line indicates the observation at $z = 9.1$ (Universe age of $\sim$550 Myr). Each component is described using age and SFR parameters. For simplicity, both components have a common dust attenuation, A_v, and metallicity, Z.



Extended Data Fig. 6 | Comparisons of constant SFH models and observational constraints. a–c, χ_ν^2 (**a**), IRAC colour (**b**) and dust emission (**c**), plotted against stellar age. All model grids are shown with grey dots, and the best-fit models at given stellar ages are indicated with red circles. In **b**, the black horizontal dashed line indicates the observed value and the yellow shaded region its 1σ uncertainty. In **c**, the black horizontal dashed line refers to the 2σ upper limit. **d–f**, Best-fit SEDs at ages of 1 Myr, 10 Myr and 100 Myr indicated by the blue squares in **a**. The insets show the flux density of the dust continuum (in μJy) and the [O III] 88-μm flux (in 10^{-18} erg cm^{-2} s^{-1}). In **d** we demonstrate that a strong nebular continuum plus [O II] 3,727-Å emission counteracts intense Hβ plus [O III] 4,959-Å emission, producing an ‘inverse Balmer break’ for very young metal-poor cases.



Extended Data Fig. 7 | Evolution of the ultraviolet luminosity of MACS1149-JD1 as a function of redshift. For each redshift bin ($\Delta z = 1$), we extrapolated the magnitude by assuming a constant SFR over the redshift interval (blue curve). We over-plotted the sensitivity of the NIRCam filters (pink and grey) covering the 1,500-Å rest frame (10σ in ~20 min) and the NIRSpec sensitivity (dashed black line) at the same wavelength (10σ in 3 h).

Extended Data Table 1 | Upper limits on the infrared luminosity and dust mass

T_d (K)	β_d	L_{TIR} ($10^9 L_\odot$)	M_d ($10^5 M_\odot$)
30	1.5	$< 6.0 \times (10/\mu)$	$< 20.0 \times (10/\mu)$
40	1.5	$< 7.7 \times (10/\mu)$	$< 5.3 \times (10/\mu)$
50	1.5	$< 11.4 \times (10/\mu)$	$< 2.3 \times (10/\mu)$
60	1.5	$< 17.3 \times (10/\mu)$	$< 1.3 \times (10/\mu)$

Upper limits represent the 3σ limits. The total luminosity, L_{TIR}, is estimated by integrating the modified black-body radiation at 8–1,000 μm. We use dust temperatures, T_d, ranging from 30 K to 60 K and an emissivity index of $\beta_d = 1.5$. The dust mass, M_d, is estimated using a dust-mass absorption coefficient of $\kappa = \kappa_0(\nu/\nu_0)^{\beta_d}$, where $\kappa_0 = 10\,\mathrm{cm}^2\,\mathrm{g}^{-1}$ at 250 μm (ref. [12]).

Extended Data Table 2 | Summary of our SED parameters

Single stellar component	
Star formation history	Exponential (τ = ±0.03, 0.06, 0.1, 0.3, 0.6, 1.0 and 10 Gyr)
Age [Myr]	0.1 – 550 (age of the Universe)
Dust A_v [mag]	0 – 1
Metallicity	Z = 0.0001, 0.0004, 0.004, 0.008 and 0.02
Star formation rate [$M_\odot$ yr^{-1}]	arbitrary (amplitude)
Two stellar components	**(Constant + Shut down)**
Age1 [Myr]	0.1 – 550 (age of the Universe)
Age2 [Myr]	0.1 – 550 (age of the Universe)
Common Dust $A_v1 = A_v2$ [mag]	0 – 1
Common Metallicity $Z1 = Z2$	Z = 0.0001, 0.0004, 0.004, 0.008 and 0.02
Total star formation rate [$M_\odot$ yr^{-1}]	arbitrary (amplitude)
SFR ratio $f_{\mathrm{SFR0}} \equiv (\mathrm{SFR2/SFR1})_{z=9.1}$	$10^{-4} - 10^{4}$

In single-component models, three SFHs with constant, exponentially declining and exponentially rising SFRs are considered. In two-components models, constant SFHs with star formation durations of $\tau = 10$ Myr, 10 Myr, 200 Myr, 300 Myr and 400 Myr are used for the old component (see Extended Data Fig. 4). The parameter f_{SFR0} defines the SFR ratio of the two components at the observation ($z = 9.1$). In all models, dust attenuation steps are $\Delta A_V = 0.1$ and age intervals follow the original 36 steps in GALEXEV[48], at 0.1–550 Myr. The SFR ratio steps are $\Delta\log(f_{\mathrm{SFR0}}) = 0.1$. The metallicity value of 0.02 corresponds to the solar metallicity.

Extended Data Table 3 | Summary of SED fit results of our fiducial model

Parameters	Values
χ^2	5.52
ν	2
χ_ν^2	2.76
Age$_{\mathrm{old}}$ [Myr]	290^{+190}_{-120}
Age$_{\mathrm{young}}$ [Myr]	3^{+2}_{-1}
Metallicity	$0.004^{+0.004}_{-0.004}$
A_{v} [mag]	$0.0^{+0.1}_{-0.0}$
Stellar mass [$10^9\ M_\odot$]	$1.1^{+0.5}_{-0.2} \times (10/\mu)$
Star formation rate [$M_\odot\ \mathrm{yr}^{-1}$]	$4.2^{+0.8}_{-1.1} \times (10/\mu)$
f_{SFR0}	4^{+76}_{-3}
$N_{ion,\ max}$ [10^{70}]	$\simeq 1 \times (10/\mu)$
$R_{s,max}$ [Mpc]	$\simeq 0.7 \times (10/\mu)^{1/3}$
Velocity shift$_{\mathrm{max}}$ [km s^{-1}]	$\simeq 900 \times (10/\mu)^{1/3}$
t_{inact} [Myr]	180^{+20}_{-60}
$z_{\mathrm{f,old}}$	$15.4^{+2.7}_{-2.0}$

Summary of the best-fit results obtained with our fiducial SED model. The uncertainties reflect the standard deviation due to the choice of a model with $\tau = 10$ Myr or $\tau = 200$ Myr. The best-fit SED is shown in Fig. 2, and its SFH is shown in red in Fig. 3. $N_{\mathrm{ion,max}}$, $R_{\mathrm{s,max}}$ and 'velocity shift$_{\mathrm{max}}$' denote the maximum values of the number of ionizing photons, the Strömgren radius and the shift of the Lyα damping wing, respectively. t_{inact} is the duration of the inactive phase and $z_{\mathrm{f,old}}$ is the formation redshift of the old component.

LETTER

https://doi.org/10.1038/s41586-018-0060-z

The pressure distribution inside the proton

V. D. Burkert[1]*, L. Elouadrhiri[1] & F. X. Girod[1]

The proton, one of the components of atomic nuclei, is composed of fundamental particles called quarks and gluons. Gluons are the carriers of the force that binds quarks together, and free quarks are never found in isolation—that is, they are confined within the composite particles in which they reside. The origin of quark confinement is one of the most important questions in modern particle and nuclear physics because confinement is at the core of what makes the proton a stable particle and thus provides stability to the Universe. The internal quark structure of the proton is revealed by deeply virtual Compton scattering[1,2], a process in which electrons are scattered off quarks inside the protons, which subsequently emit high-energy photons, which are detected in coincidence with the scattered electrons and recoil protons. Here we report a measurement of the pressure distribution experienced by the quarks in the proton. We find a strong repulsive pressure near the centre of the proton (up to 0.6 femtometres) and a binding pressure at greater distances. The average peak pressure near the centre is about 10^{35} pascals, which exceeds the pressure estimated for the most densely packed known objects in the Universe, neutron stars[3]. This work opens up a new area of research on the fundamental gravitational properties of protons, neutrons and nuclei, which can provide access to their physical radii, the internal shear forces acting on the quarks and their pressure distributions.

The basic mechanical properties of the proton are encoded in the gravitational form factors (GFFs) of the energy–momentum tensor[1,4,5]. Graviton–proton scattering is the only known process that can be used to directly measure these form factors[4,6], whereas generalized parton distributions[2,7,8] enable indirect access to the basic mechanical properties of the proton[2].

A direct determination of the quark pressure distribution in the proton (Fig. 1) requires measurements of the proton matrix element of the energy–momentum tensor[9]. This matrix element contains three scalar GFFs that depend on the four-momentum transfer t to the proton. One of these GFFs, $d_1(t)$, encodes the shear forces and pressure distribution on the quarks in the proton, and the other two, $M_2(t)$ and $J(t)$, encode the mass and angular momentum distributions. Experimental information on these form factors is essential to gain insight into the dynamics of the fundamental constituents of the proton. The framework of generalized parton distributions (GPDs)[2,7,8] has provided a way to obtain information on $d_1(t)$ from experiments. The most effective way to access GPDs experimentally is deeply virtual Compton scattering (DVCS)[1,2], where high-energy electrons (e) are scattered from the protons (p) in liquid hydrogen as $e\,p \to e'\,p'\,\gamma$, and the scattered electron (e'), proton (p') and photon (γ) are detected in coincidence. In this process, the quark structure is probed with high-energy virtual photons that are exchanged between the scattered electron and the proton, and the emitted (real) photon controls the momentum transfer t to the proton, while leaving the proton intact. Recently, methods have been developed to extract information about the GPDs and the related Compton form factors (CFFs) from DVCS data[10–13].

To determine the pressure distribution in the proton from the experimental data, we follow the steps that we briefly describe here. We note that the GPDs, CFFs and GFFs apply only to quarks, not to gluons.
(1) We begin with the sum rules that relate the Mellin moments of the GPDs to the GFFs[1].
(2) We then define the complex CFF, $\mathcal{H}$, which is directly related to the experimental observables describing the DVCS process, that is, the differential cross-section and the beam-spin asymmetry.
(3) The real and imaginary parts of $\mathcal{H}$ can be related through a dispersion relation[14–16] at fixed t, where the term $D(t)$, or D-term, appears as a subtraction term[17].
(4) We derive $d_1(t)$ from the expansion of $D(t)$ in the Gegenbauer polynomials of ξ, the momentum transfer to the struck quark.
(5) We apply fits to the data and extract $D(t)$ and $d_1(t)$.
(6) Then, we determine the pressure distribution from the relation between $d_1(t)$ and the pressure $p(r)$, where r is the radial distance from the proton's centre, through the Bessel integral.

The sum rules that relate the second Mellin moments of the chiral-even GPDs to the GFFs are[1]:

$$\int x\,[H(x,\xi,t)+E(x,\xi,t)]\,\mathrm{d}x = 2J(t)$$

$$\int xH(x,\xi,t)\mathrm{d}x = M_2(t)+\frac{4}{5}\xi^2 d_1(t)$$

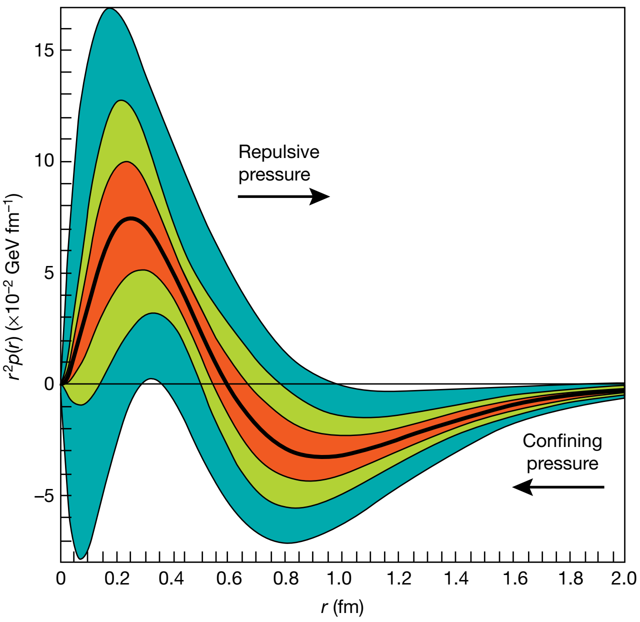


Fig. 1 | Radial pressure distribution in the proton. The graph shows the pressure distribution $r^2p(r)$ that results from the interactions of the quarks in the proton versus the radial distance r from the centre of the proton. The thick black line corresponds to the pressure extracted from the D-term parameters fitted to published data[22] measured at 6 GeV. The corresponding estimated uncertainties are displayed as the light-green shaded area shown. The blue area represents the uncertainties from all the data that were available before the 6-GeV experiment, and the red shaded area shows projected results from future experiments at 12 GeV that will be performed with the upgraded experimental apparatus[30]. Uncertainties represent one standard deviation.

---

[1] Thomas Jefferson National Accelerator Facility, Newport News, VA, USA. *e-mail: burkert@jlab.org

where H and E are the GPDs corresponding to the nucleon helicity-conserving and helicity-flip processes, respectively; x and ξ are the average and the transferred quark momentum fractions, respectively; and $M_2(t)$, $J(t)$ and $d_1(t)$ correspond to the time–time, time–space and space–space components of the energy–momentum tensor, respectively. We have some constraints on $M_2(t)$ and $J(t)$ from theory and previous work; namely, at $t=0$ they are fixed to the proton's mass and spin, respectively. By contrast, almost nothing is known about the equally fundamental quantity $d_1(t)$. As the GFF $d_1(t)$ encodes the shear forces on the quarks and the pressure distribution in the proton, we can expect the existence of a zero-sum rule ensuring that the total pressure and forces vanish, thus preserving the stability of the dynamics. The observables are parameterized by the CFFs, which for the GPD H are the real quantities $\mathrm{Re}\mathcal{H}$ and $\mathrm{Im}\mathcal{H}$ defined by:

$$\mathrm{Re}\mathcal{H}(\xi,t)+i\mathrm{Im}\mathcal{H}(\xi,t)=\int_{-1}^{1}\mathrm{d}x\left(\frac{1}{\xi-x-\mathrm{i}\epsilon}-\frac{1}{\xi+x-\mathrm{i}\epsilon}\right)H(x,\xi,t) \quad (1)$$

The average quark momentum fraction x is not observable in this process; it is integrated over with the inverse of the quark momentum fractions shifted from the real axis by an infinitesimal positive number ϵ. The analytical properties of the amplitude $\mathcal{H}$ (as a complex function) in the leading-order approximation lead to the dispersion relation:

$$\mathrm{Re}\mathcal{H}(\xi,t)\overset{\mathrm{lo}}{=}D(t)+\mathcal{P}\int_{-1}^{1}\mathrm{d}x\left(\frac{1}{\xi-x}-\frac{1}{\xi+x}\right)\mathrm{Im}\mathcal{H}(x,t)$$

where $\mathcal{P}$ is the principal part of the integral and the subtraction constant is the D-term. This approach allows us to replace one unknown complex CFF in the dispersion relation with its imaginary part and the D-term[18,19].

We derive $d_1(t)$ as the first coefficient in the Gegenbauer expansion of the D-term. Here, we truncate this expansion to $d_1(t)$ only.

$$D(t)=\frac{1}{2}\int_{-1}^{1}\frac{D(z,t)}{1-z}\mathrm{d}z$$

with

$$D(z,t)=(1-z^2)\,[d_1(t)C_1^{3/2}(z)+\cdots]$$

and

$$-1<z=\frac{x}{\xi}<1$$

C_1 is the first coefficient in the Gegenbauer expansion, and z is the ratio of the quark momentum fraction to the longitudinal momentum transfer to the quark. Our starting points in the analysis are the global fits presented in refs [10,11], referred to as Kumerički–Müller parameterization. The imaginary part of the amplitude is calculated from a parameterization of the GPDs along the diagonal $x=\xi$. The real part of the amplitude is then reconstructed by assuming leading-order dominance and applying the dispersion relation. The ξ-dependence of the D-term is completely generated by the Gegenbauer expansion and restricted to the $d_1(t)$ term only. Finally, the momentum transfer dependence of the $d_1(t)$ term is given as a function of the three parameters $d_1(0)$, M and α:

$$d_1(t)=d_1(0)\left(1-\frac{t}{M^2}\right)^{-\alpha}$$

where M and α are parameters adjusted to the data, and the chosen form of $d_1(t)$ with $\alpha=3$ is consistent with the asymptotic behaviour required by the dimensional counting rules in QCD[20]. We adjust and fix the central values of the model parameters to DVCS data obtained at an electron beam energy of 6 GeV[21,22] which include unpolarized and polarized beam cross-sections over a wide phase space in the valence region and support the model that suggests that the GPD H largely dominates these observables. The fit to the beam-spin asymmetries is shown in Fig. 2, and that to the unpolarized cross-sections in Fig. 3. A fit to $d_1(t)$ is provided in Fig. 4; the data points correspond to the values extracted from the fit in Fig. 3. The analysis of the experimental results shows that $d_1(0)$ has a negative sign, consistent with several theoretical studies[16,23,24]. The fit results give $d_1(0)=-2.04$ with statistical uncertainty 0.14 and systematic uncertainty 0.33, where the uncertainties are

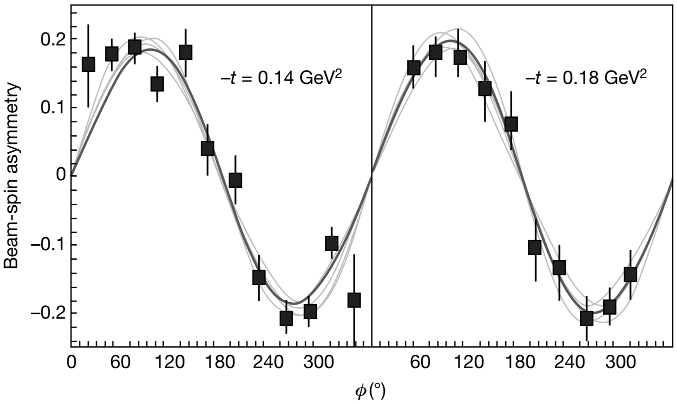


Fig. 2 | Fits to the beam-spin asymmetry results. The experimental results (squares) are shown as a function of the angle ϕ between the electron and recoil scattering planes. The sine-like behaviour of the data is due to the interference of DVCS and Bethe–Heitler processes. The fit is shown as the thick dark curve. The thin light-grey curves correspond to variations of the fitting parameters to the imaginary part of the CFF $\mathcal{H}$, as calculated from a Kumerički–Müller parameterization[10] adjusted to the data. Uncertainties represent one standard deviation.

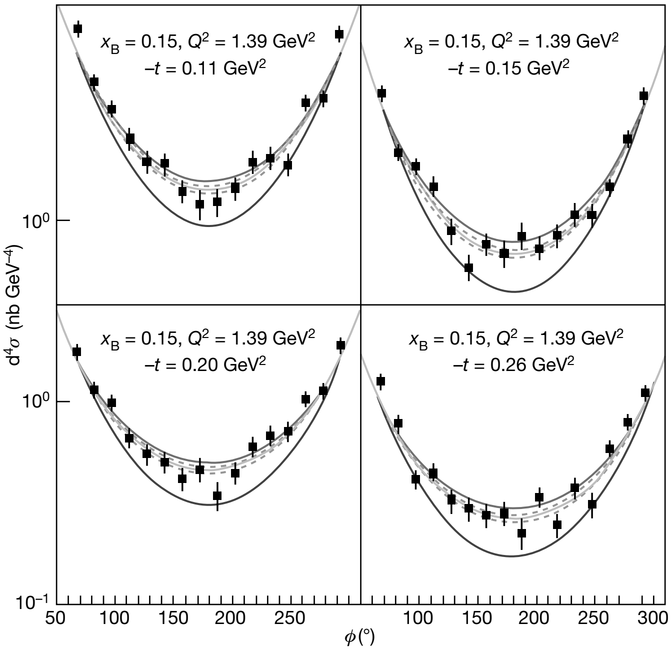


Fig. 3 | Examples of fits to unpolarized cross-sections. Experimental DVCS cross-sections[22] (black squares) are shown as a function of ϕ at fixed values of x_{B} and Q^2. Q^2 is the virtuality of the exchanged photon, and x_{B} is the momentum fraction of the struck quark. The light-grey curves show the results of the local fits with the parameter $d_1(t)$ at fixed $-t$, including one standard deviation. The upper, dark-grey curves are the results of global fits to the $-t$ dependence of $d_1(t)$, with the real part of the DVCS amplitude calculated from the dispersion relations and the subtraction term evaluated from the $d_1(t)$ contribution. The Bethe–Heitler contribution is shown with the lower, black curve. Its contribution to the cross-section dominates the extreme regions of ϕ. Uncertainties represent one standard deviation.

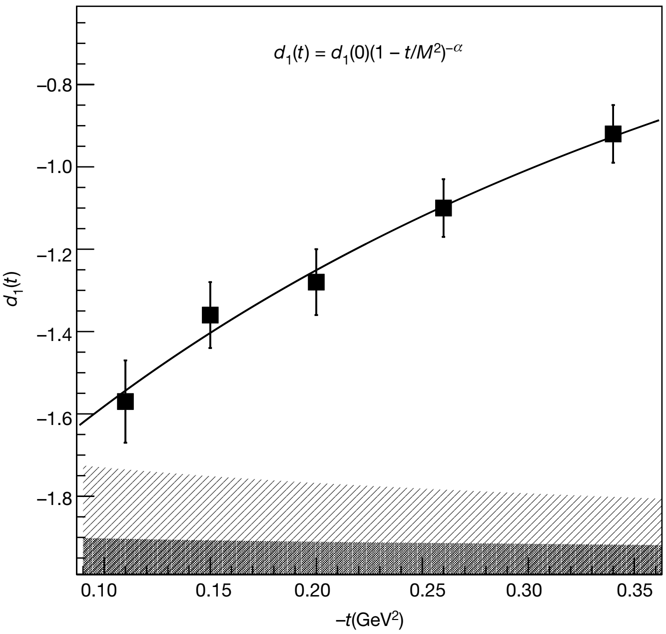


Fig. 4 | Example of a fit to gravitational form factor $d_1(t)$. The error bars are from the fit to the cross-sections at a fixed value of $-t$. The light-shaded area at the bottom corresponds to the uncertainties from the extension of the fit into regions without data and coincides with the light-green shaded area in Fig. 1. The dark-shaded area corresponds to the projected uncertainties for a future experiment[30] at 12 GeV, as shown by the red-shaded area in Fig. 1. Uncertainties represent one standard deviation.

given as one standard deviation. The negative sign of $d_1(0)$ found in this analysis seems deeply rooted in the spontaneous breakdown of chiral symmetry[25], which is a consequence of the transition of the micro-second-old Universe from a state of de-confined quarks and gluons to a state of confined quarks in stable protons. It is thus intimately connected with the stability of the proton[24] and of the visible Universe.

We can relate $d_1(t)$ to the pressure distribution via the spherical Bessel integral:

$$d_1(t) \propto \int \frac{j_0(r\sqrt{-t})}{2t} p(r)\,\mathrm{d}^3 r$$

where j_0 is the first spherical Bessel function. Our results of the quark pressure distribution in the proton are illustrated in Fig. 1. The thick black line corresponds to the pressure distribution $r^2 p(r)$, as extracted from the D-term parameters that are fitted to the published data[22] acquired at 6 GeV. The estimated uncertainties are displayed as the light-green shaded area. The red-shaded area represents projected results from future experiments at higher energy. The distribution has a positive core and a negative tail of the $r^2 p(r)$ distribution as a function of r, with a zero crossing near $r = 0.6$ fm. The regions where repulsive and binding pressures dominate are separated in radial space, with the repulsive distribution peaking near $r = 0.25$ fm, and the maximum of the negative pressure that is responsible for the binding occurring near $r = 0.8$ fm.

The outer, blue-shaded area in Fig. 1 corresponds to the D-term uncertainties obtained in the global fit results from previous research[10,11]. This area has a shape similar to the light-green area, confirming the robustness of the analysis procedure used to extract the D-term. The pressure $p(r)$ must satisfy the stability condition:

$$\int_0^\infty r^2 p(r)\,\mathrm{d}r = 0$$

which is satisfied within the uncertainties of our analysis. The shape of the radial pressure distribution resembles closely that obtained using the chiral quark–soliton model[24], in which the proton is modelled as a chiral soliton whose constituent quarks are bound by a self-consistent pion field. This agreement suggests that the pion field is appropriate for the description of the proton as a bound state of quarks.

Other applications of the GFFs of the energy–momentum tensor include the description of nucleons in the nuclear medium[23,26,27], excited baryon states (such as the $\Delta(1232)$ resonance[28]) and point-like and composed spin-0 particles[29].

Future precision experiments are expected to provide substantially more DVCS data[30] and enable the mapping of $d_1(t)$ in much finer steps and in a much larger $-t$ range, which will reduce the systematic uncertainties, as indicated by the red-shaded area in Fig. 1. We also expect that this work will motivate new theoretical efforts to understand the fundamental characteristics of the stability of the proton from first principles. Our results may serve as a benchmark for the assessment of theoretical models, including lattice quantum chromodynamics models.

Online content

Any Methods, including any statements of data availability and Nature Research reporting summaries, along with any additional references and Source Data files, are available in the online version of the paper at https://doi.org/10.1038/s41586-018-0060-z.

Received: 25 August 2017; Accepted: 16 February 2018;
Published online 16 May 2018.

1. Ji, X. D. Deeply virtual Compton scattering. *Phys. Rev. D* **55**, 7114–7125 (1997).
2. Ji, X. D. Gauge-invariant decomposition of nucleon spin. *Phys. Rev. Lett.* **78**, 610–613 (1997).
3. Ozel, F. & Freire, P. Masses, radii, and equation of state of neutron stars. *Annu. Rev. Astron. Astrophys.* **54**, 401–440 (2016).
4. Pagels, H. Energy–momentum structure form factors of particles. *Phys. Rev.* **144**, 1250–1260 (1966).
5. Teryaev, O. V. Gravitational form factors and nucleon spin structure. *Front. Phys.* **11**, 111207 (2016).
6. Belitsky, A. V. & Radyushkin, A. V. Unraveling hadron structure with generalized parton distributions. *Phys. Rep.* **418**, 1–387 (2005).
7. Müller, D., Robaschik, D., Geyer, D., Dittes, F. M. & Horejši, J. Wave functions, evolution equations and evolution kernels from light-ray operators of QCD. *Fortschr. Phys.* **42**, 101–141 (1994).
8. Radyushkin, A. V. Scaling limit of deeply virtual Compton scattering. *Phys. Lett. B* **380**, 417–425 (1996).
9. Polyakov, M. V. Generalized parton distributions and strong forces inside nucleons and nuclei. *Phys. Lett. B* **555**, 57–62 (2003).
10. Kumerički, K. & Müller, D. Deeply virtual Compton scattering at small x_B and the access to the GPD H. *Nucl. Phys. B* **841**, 1–58 (2010).
11. Müller, D., Lautenschlager, T., Passek-Kumericki, K. & Schaefer, A. Towards a fitting procedure to deeply virtual meson production – the next-to-leading order case. *Nucl. Phys. B* **884**, 438–546 (2014).
12. Guidal, M., Moutarde, H. & Vanderhaeghen, M. Generalized parton distributions in the valence region from deeply virtual Compton scattering. *Rep. Prog. Phys.* **76**, 066202 (2013).
13. Kumerički, K., Liuti, S. & Moutarde, H. GPD phenomenology and DVCS fitting: entering the high-precision era. *Eur. Phys. J. A* **52**, 157 (2016).
14. Diehl, M. & Ivanov, D. Y. Dispersion representations for hard exclusive processes: beyond the Born approximation. *Eur. Phys. J. C* **52**, 919–932 (2007).
15. Anikin, I. V. & Teryaev, O. V. Dispersion relations and QCD factorization in hard reactions. *Fizika B* **17**, 151–158 (2008).
16. Pasquini, B., Polyakov, M. V. & Vanderhaeghen, M. Dispersive evaluation of the D-term form factor in deeply virtual Compton scattering. *Phys. Lett. B* **739**, 133–138 (2014).
17. Polyakov, M. V. & Weiss, C. Skewed and double distributions in the pion and the nucleon. *Phys. Rev. D* **60**, 114017 (1999).
18. Radyushkin, A. V. Modeling nucleon generalized parton distributions. *Phys. Rev. D* **87**, 096017 (2013).
19. Radyushkin, A. V. Sum rules for nucleon generalized parton distributions and border function formulation. *Phys. Rev. D* **88**, 056010 (2013).
20. Lepage, G. P. & Brodsky, S. J. Exclusive processes in perturbative quantum chromodynamics. *Phys. Rev. D* **22**, 2157–2198 (1980).
21. CLAS Collaboration. Measurement of deeply virtual Compton scattering beam-spin asymmetries. *Phys. Rev. Lett.* **100**, 162002 (2008).
22. CLAS Collaboration. Cross sections for the exclusive photon electroproduction on the proton and generalized parton distributions. *Phys. Rev. Lett.* **115**, 212003 (2015).
23. Kim, H. C., Schweitzer, P. & Yakhshiev, U. Energy–momentum tensor form factors of the nucleon in nuclear matter. *Phys. Lett. B* **718**, 625–631 (2012).
24. Goeke, K. et al. Nucleon form factors of the energy–momentum tensor in the chiral quark–soliton model. *Phys. Rev. D* **75**, 094021 (2007).

25. Kivel, N., Polyakov, M. V. & Vanderhaeghen, M. Deeply virtual Compton scattering on the nucleon: study of the twist-3 effects. *Phys. Rev. D* **63**, 114014 (2001).
26. Eides, M. I., Petrov, V. Y. & Polyakov, M. V. Narrow nucleon–$\psi(2S)$ bound state and LHCb pentaquarks. *Phys. Rev. D* **93**, 054039 (2016).
27. Jung, J. H., Yakhshiev, U., Kim, H. C. & Schweitzer, P. In-medium modified energy–momentum tensor form factors of the nucleon within the framework of a π–ρ–ω soliton model. *Phys. Rev. D* **89**, 114021 (2014).
28. Perevalova, I. A., Polyakov, M. V. & Schweitzer, P. LHCb pentaquarks as a baryon–ψ(2S) bound state: prediction of isospin-3/2 pentaquarks with hidden charm. *Phys. Rev. D* **94**, 054024 (2016).
29. Hudson, J. & Schweitzer, P. *D* term and the structure of pointlike and composed spin-0 particles. *Phys. Rev. D* **96**, 114013 (2017).
30. CLAS Collaboration *Deeply Virtual Compton Scattering with CLAS12 at 6.6 GeV and 8.8 GeV*. Proposal E12–16–010B (Jefferson Lab PAC44, 2016); https://www.jlab.org/exp_prog/proposals/16/PR12-16-010B.pdf.

Acknowledgements We thank M. Polyakov and P. Schweitzer for discussions on the subject matter of this work. We are indebted to G. D. Cates for reading the manuscript and suggesting text improvements. This material is based on work supported by the US Department of Energy, Office of Science, Office of Nuclear Physics under contract DE-AC05-06OR23177.

Reviewer information Nature thanks G. D. Cates and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions V.D.B. wrote the manuscript initiated the analysis, and coordinated and oversaw the research and the validation of the results. L.E. worked with theory experts on the feasibility of the present analysis to develop a procedure connecting the data to the pressure measurements and worked on the analysis of the DVCS beam spin asymmetry. F.X.G. provided the detailed analysis of beam asymmetry and the cross-section measurements, performed the fits to extract the D-term and used the dispersion relation analysis to relate the results to the confinement form factor $d_1(t)$.

Competing interests The authors declare no competing interests.

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0060-z.

Correspondence and requests for materials should be addressed to V.D.B.

METHODS

Here we provide additional details about the method of extraction of the CFFs using dispersion relations, from which the D-term and $d_1(t)$ are computed. In the first step, the beam-spin asymmetries and unpolarized cross-sections are fitted simultaneously, and the CFFs are obtained locally as the real and imaginary parts of the amplitude at one particular kinematical point. In the second step, the CFFs are fitted globally with the Kumerički–Müller parameterization[10,11], the real part of the amplitude is calculated by the dispersion integral from the imaginary part of the amplitude, and the D-term is derived as a subtraction, as described in the text.

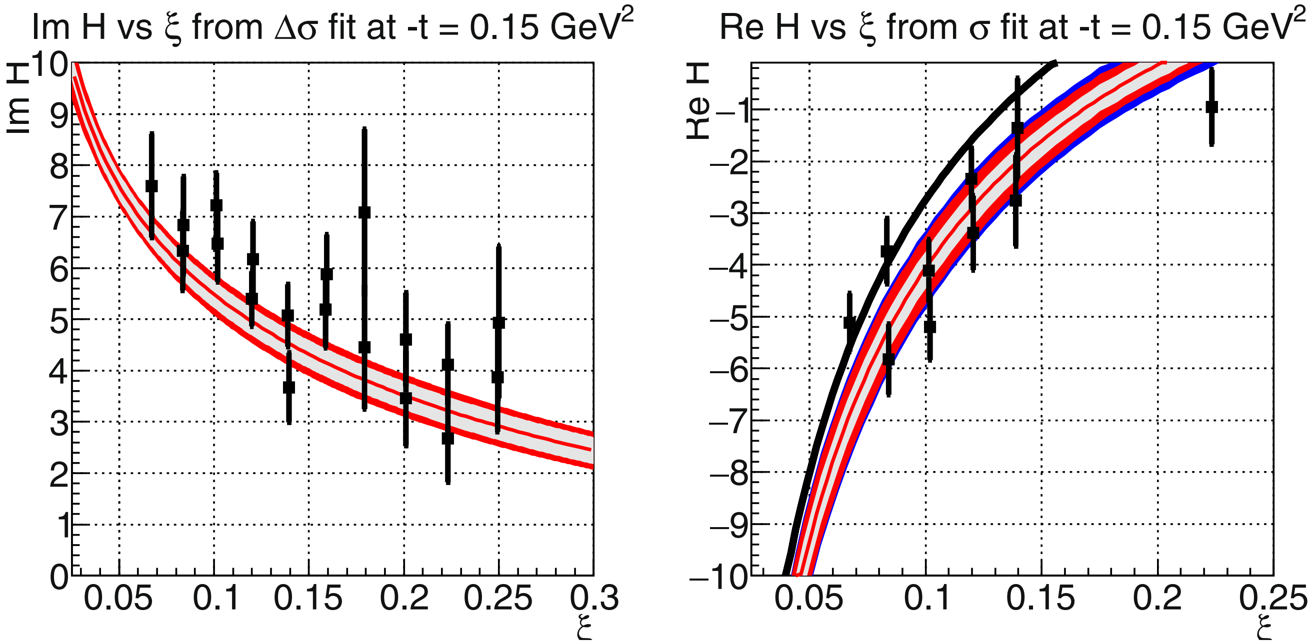
To illustrate the robustness of our results, in Extended Data Fig. 1 we show the locally extracted CFFs Im$\mathcal{H}$ and Re$\mathcal{H}$ separately as a function of ξ, for a fixed t value. These locally extracted amplitudes (black squares) are compared to the global parameterization (central red curve, with the grey bands showing the estimated uncertainties). The left panel shows the imaginary part of the amplitude $\mathcal{H}$, which is mostly sensitive to beam-spin asymmetries, together with the amplitude parameterization as a function of ξ at $-t=0.13\,\mathrm{GeV}^2$. The interpretation of the error bands as separate contributions from different uncertainties in the parameterization, after adjustment to our data, are discussed in more detail below. The right panel shows the real part of the amplitude, which is mostly sensitive to the unpolarized cross-sections. The different numbers of points in the polarized and unpolarized cross-sections come from the real and imaginary parts of the amplitude being sensitive to different ranges in the azimuthal angle ϕ. The real part of the amplitude is more sensitive in the central range around $\phi=180°$, where the statistics and the acceptances are reduced—as a consequence, Re$\mathcal{H}$ cannot be determined when ξ increases.

The central red curve in each panel is the result of our global fit. The grey band between the thick red lines shows the uncertainties of the contribution from the other CFFs, on which our data do not provide strong constraints. The red band shows the total error of the global fit to the imaginary part of the amplitude, including fitting errors of the model parameters, in addition to the above uncertainties. In our framework, these grey and red bands are propagated through the dispersion integral as direct uncertainties in the real part of the amplitude. The additional blue band in the right panel illustrates the total error of the real part of the amplitude, including that of the extracted D-term. We note that the uncertainties in the D-term increase with $-t$, as reflected by the larger uncertainties in the spatial distribution at short distances, even after r^2 weighting.

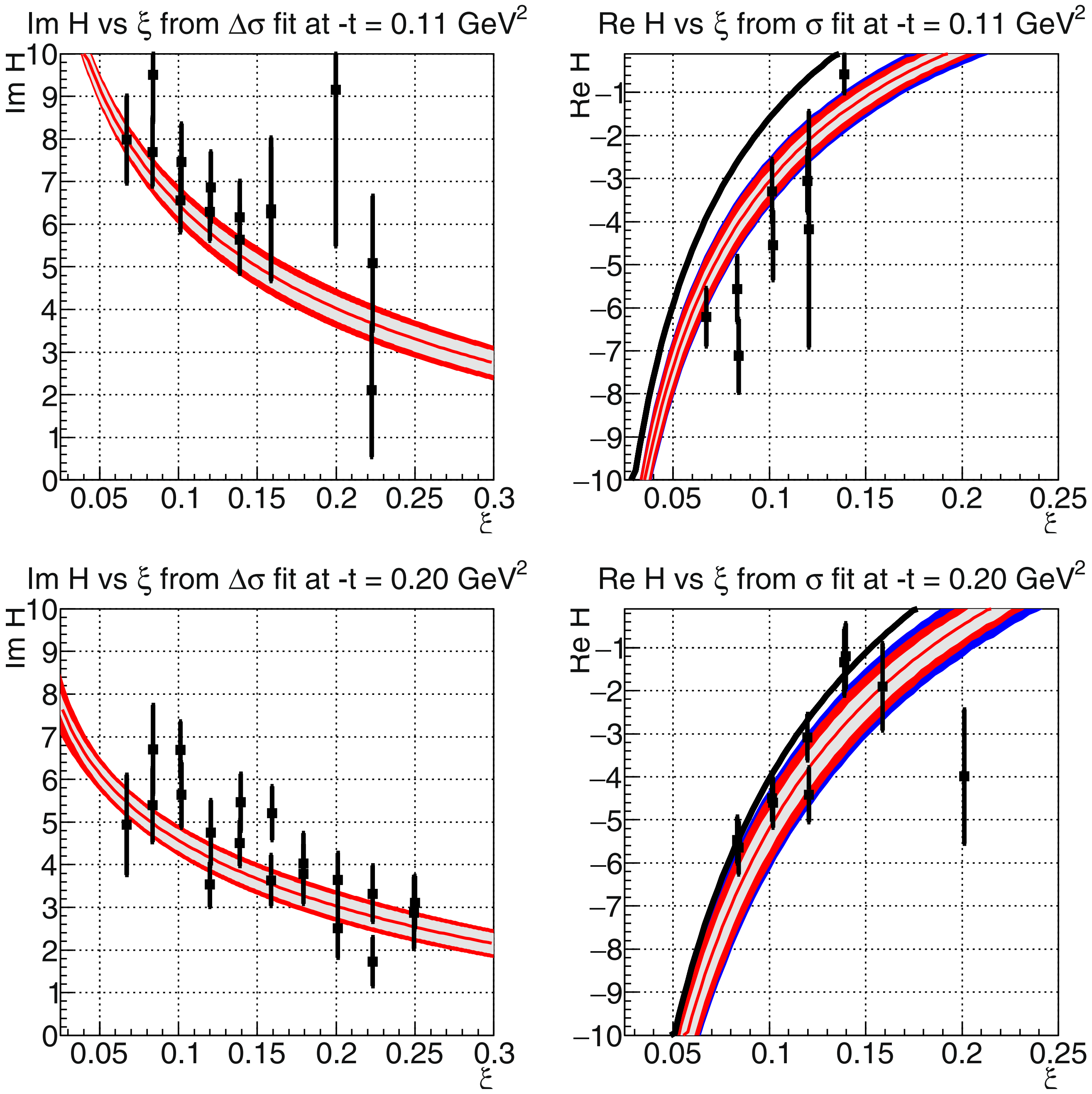
For completeness, we show in Extended Data Fig. 2 and Extended Data Fig. 3 the real and imaginary parts of the amplitude for the other four $-t$ values that are shown in Fig. 4 as a function of ξ. Our data show a behaviour similar to the Kumerički–Müller parameterization at the lower $-t$ value, but begin to depart from the calculations at higher $-t$ values. The assumption $D(t)=0$, shown by the black line, is inconsistent with our full results. This confirms that finite values of $D(t)$ are required and again demonstrates the robustness of our results shown in Fig. 1.

Data availability. The data used in this work are publicly available at http://clasweb.jlab.org/physicsdb/intro.html.

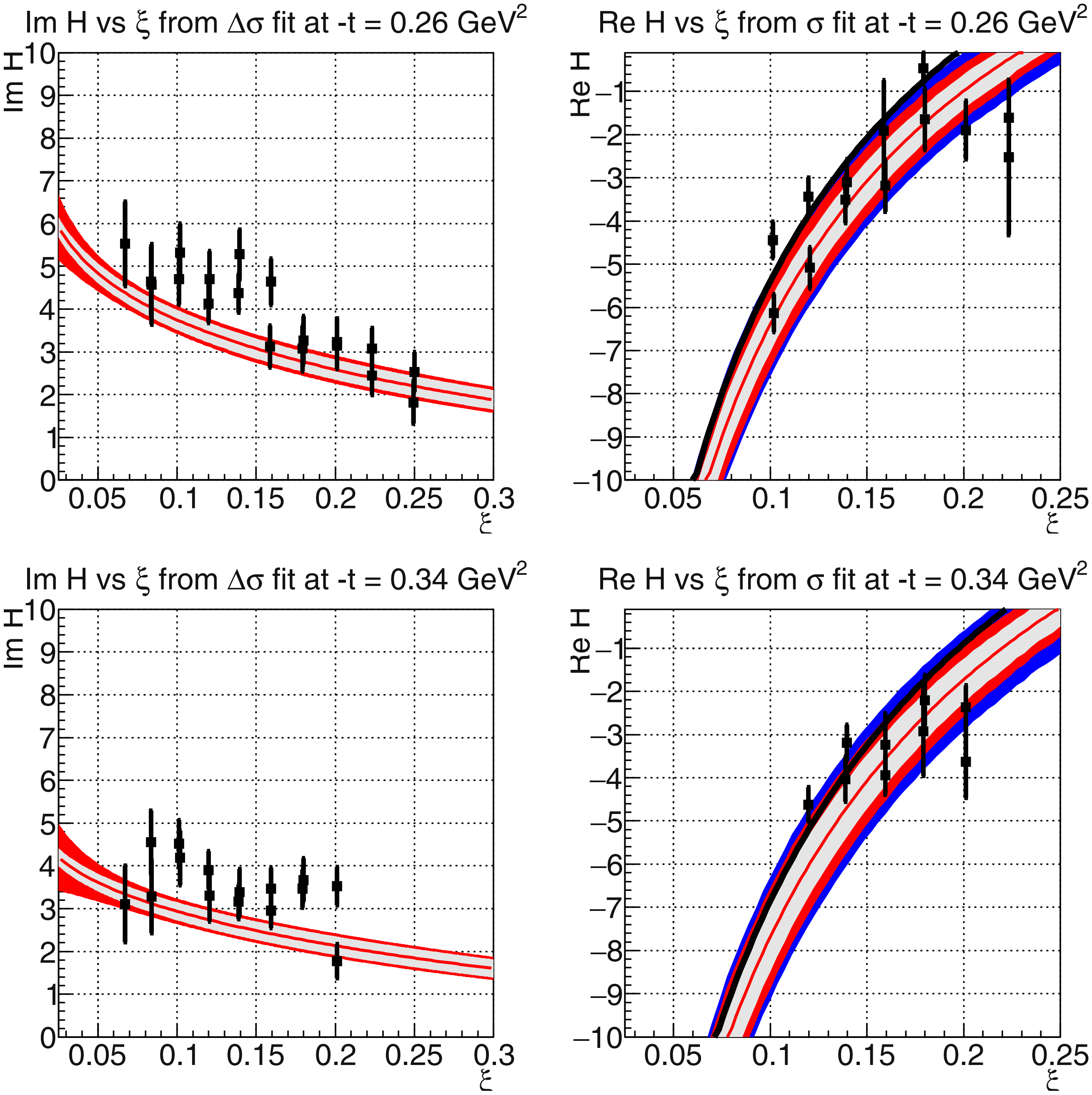
Code availability. The code used in the data analysis is available at https://userweb.jlab.org/~fxgirod/code_paperFeb18. The data reduction program uses the standard C++ ROOT framework, which was developed at CERN and is freely available under a GNU Lesser General Public License at https://root.cern.ch. The source code to reproduce the figures is freely available at https://userweb.jlab.org/~fxgirod/code_paperFeb18/.



Extended Data Fig. 1 | Compton form factor $\mathcal{H}$ at $-t = 0.15\,\mathrm{GeV}^2$. The graphs show the imaginary (left) and real (right) parts of $\mathcal{H}$ versus the momentum transfer to the quark, ξ. The inner red curve shows the result of our global fit. The grey band shows the estimated uncertainties from the contributions of other CFFs, the outer red band shows the total uncertainties of the imaginary part of the amplitude, and the outer blue band (right) includes the uncertainties related to the D-term. All uncertainties represent one standard deviation.



Extended Data Fig. 2 | Compton form factor $\mathcal{H}$ at $-t=0.11$ GeV^2 and at $-t=0.20$ GeV^2. The imaginary (left) and real (right) parts of $\mathcal{H}$ are shown as a function of ξ. The red curve and the grey, red and blue bands are as in Extended Data Fig. 1. All uncertainties represent one standard deviation.



Extended Data Fig. 3 | Compton form factor $\mathcal{H}$ at $-t = 0.26\,\mathrm{GeV}^2$ and at $-t = 0.34\,\mathrm{GeV}^2$. The imaginary (left) and real (right) parts of $\mathcal{H}$ are shown as a function of ξ. The red curve and the grey, red and blue bands are as in Extended Data Fig. 1. All uncertainties represent one standard deviation.

LETTER

https://doi.org/10.1038/s41586-018-0107-1

Dynamic band-structure tuning of graphene moiré superlattices with pressure

Matthew Yankowitz[1], Jeil Jung[2], Evan Laksono[3,4], Nicolas Leconte[2], Bheema L. Chittari[2], K. Watanabe[5], T. Taniguchi[5], Shaffique Adam[3,4,6], David Graf[7] & Cory R. Dean[1]*

Heterostructures can be assembled from atomically thin materials by combining a wide range of available van der Waals crystals, providing exciting possibilities for designer electronics[1]. In many cases, beyond simply realizing new material combinations, interlayer interactions lead to emergent electronic properties that are fundamentally distinct from those of the constituent layers[2]. A critical parameter in these structures is the interlayer coupling strength, but this is often not easy to determine and is typically considered to be a fixed property of the system. Here we demonstrate that we can controllably tune the interlayer separation in van der Waals heterostructures using hydrostatic pressure, providing a dynamic way to modify their electronic properties. In devices in which graphene is encapsulated in boron nitride and aligned with one of the encapsulating layers, we observe that increasing pressure produces a superlinear increase in the moiré-superlattice-induced bandgap—nearly doubling within the studied range—together with an increase in the capacitive gate coupling to the active channel by as much as 25 per cent. Comparison to theoretical modelling highlights the role of atomic-scale structural deformations and how this can be altered with pressure. Our results demonstrate that combining hydrostatic pressure with controlled rotational order provides opportunities for dynamic band-structure engineering in van der Waals heterostructures.

Heterostructures fabricated by mechanical assembly of atomically thin van der Waals (vdW) crystals represent an exciting new paradigm in materials design. Owing to weak interlayer bonding, two-dimensional (2D) crystals with wide-ranging characteristics and composition—such as graphene, boron nitride (BN) and the transition metal dichalcogenides—can be readily mixed and matched, without the usual interfacial constraints of conventional crystal growth[3]. Moreover, atomic-scale crystalline alignment between the layers often plays a critical role in the resulting device characteristics, leading to additional and controllable degrees of freedom. For example, electronic coupling processes that are sensitive to momentum mismatch, such as interlayer tunnelling[4] or exciton binding[5], can be sensitively tuned by varying the rotational order. For crystals with closely matched lattice constants, moiré interference at zero-angle alignment can additionally result in long-range superlattice potentials, which in turn can lead to entirely new electronic device characteristics[6–14].

BN-encapsulated graphene provides a model example of the variety of electronic properties that can be realized on demand in a single type of vdW heterostructure. At large relative twist angles, BN acts as a featureless dielectric for graphene, minimizing coupling to extrinsic disorder but otherwise remaining effectively inert[15]. However, at small twist angle, coupling to the resulting moiré superlattice (MSL) profoundly alters the graphene band structure, giving rise to secondary Dirac cones at finite energy[6,16] while also modifying the Fermi velocity near the Dirac point[16]. As a consequence, several unusual electronic properties have been observed, such as density-dependent topological valley currents at zero magnetic field[11] and the fractal Hofstadter butterfly spectrum at high field[7–9]—recently identified as hosting integer and fractional Chern insulating states[14] and charge density waves[12,17]. Additionally, the MSL opens bandgaps at both the primary and secondary Dirac points of graphene, which are of particular interest for graphene-based digital logic.

Numerous techniques have been developed to control the rotational alignment in graphene/BN and related vdW heterostructures[7,12,13,18–21]. Equally important is the spacing between the layers, which dictates the magnitude of interlayer interactions. Little experimental work has

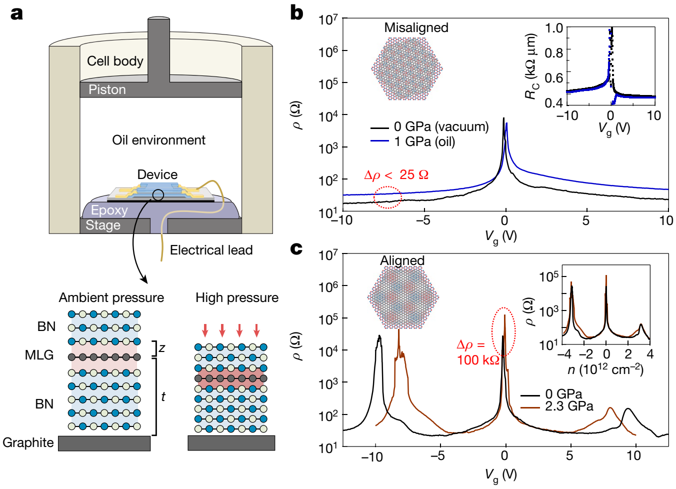


Fig. 1 | Experiment schematic and transport characterization under pressure. a, Top, cartoon of the piston-cylinder pressure cell. The sample is mounted inside a Teflon cup filled with oil. The pressure exerted on the sample is determined by the force applied to the piston. Electrical feedthroughs embedded in an epoxy junction enable in situ transport measurements. Bottom, schematic of the sample of graphene (MLG, monolayer graphene) encapsulated between two flakes of BN with a graphite gate. Neighbouring layers sit a distance z from one another. The distance between the top graphite layer and the graphene t sets the dielectric thickness. On the right, the same structure is shown under pressure, with each layer compressed towards its neighbour, decreasing both z and t and enhancing the interlayer electronic coupling (red hue). **b**, Resistivity of a misaligned device (device P2) in vacuum and at high pressure (1 GPa) in oil, exhibiting little pressure dependence. The contact resistance R_c (taken as the difference between the two- and four-terminal resistances) also exhibits virtually no dependence on pressure (inset). **c**, Resistivity of an aligned device (P1). The SDPs appear at different gate voltages as a function of pressure, but are perfectly aligned when plotted against n (inset). The PDP becomes noticeably more insulating with pressure, as highlighted in the red dashed oval. Cartoon insets in **b** and **c** show schematics of the moiré patterns that arise from the relative alignment of the graphene and BN. All measurements are taken at $B = 0$ T and $T = 2$ K.

---

[1]Department of Physics, Columbia University, New York, NY, USA. [2]Department of Physics, University of Seoul, Seoul, South Korea. [3]Centre for Advanced 2D Materials, National University of Singapore, Singapore, Singapore. [4]Department of Physics, Faculty of Science, National University of Singapore, Singapore, Singapore. [5]National Institute for Materials Science, Tsukuba, Japan. [6]Yale-NUS College, Singapore, Singapore. [7]National High Magnetic Field Laboratory, Tallahassee, FL, USA. *e-mail: cd2478@columbia.edu

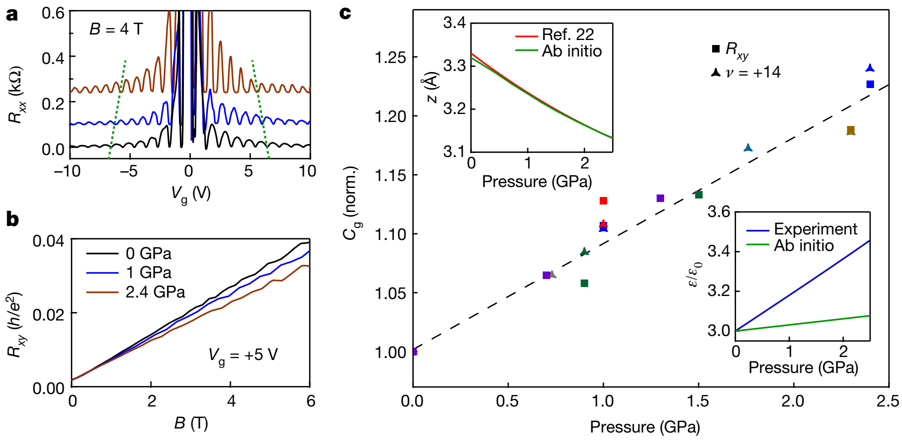


Fig. 2 | Modification of gate capacitance under pressure. a, Longitudinal resistance R_{xx} of device P2 at $B = 4$ T and $T = 2$ K. The green dotted lines track the same quantum Hall feature with pressure. Curves are vertically offset proportional to the pressure applied. **b**, R_{xy} as a function of pressure in the same device at $V_g = +5$ V. **c**, Gate capacitance C_g as a function of pressure, normalized to its value at 0 GPa for each device. Colours represent measurements on different devices. The square markers calculate C_g using n extracted from the low-field R_{xy}. The triangle markers use n extracted from the high-field dispersion of $\nu = 14$, where ν is the filling factor. Top inset, change in BN interlayer spacing per layer as a function of pressure predicted by LDA ab initio modelling (green curve). This almost exactly matches previous measurements by X-ray diffraction[22] (reproduced in the red curve). Bottom inset, he remaining increase in C_g with pressure is attributed to an enhancement of the BN dielectric constant ε, assuming $\varepsilon = 3$ at 0 GPa as taken from experiment (blue curve). The green curve shows the LDA ab initio modelling of bulk BN.

been done to characterize or control this parameter, which is often not well-known and is assumed to be invariable. Here we demonstrate that by applying hydrostatic pressure to BN-encapsulated graphene, we can decrease the interlayer spacing by more than 5%. At small rotation angles, the resulting increase in the effective MSL potential substantially modifies the electronic device characteristics. Most notably, with increasing pressure we observe a superlinear increase in the moiré-induced bandgap at the primary Dirac point (PDP)—nearly doubling over the pressure ranges studied—to yield the largest gap so far demonstrated in pristine monolayer graphene by 35%. By contrast, the gap at the secondary Dirac point (SDP) shows little change with pressure. This unexpected result provides new insight into the precise influence of the MSL on the graphene layer and suggests that in addition to electrostatic coupling, lattice-scale deformations play an important role. Our findings reveal that interlayer spacing in vdW heterostructures is an important and tunable degree of freedom that provides a new route to band-structure engineering.

Figure 1a shows a cartoon schematic of our experimental set-up. We fabricate BN-encapsulated graphene devices using the vdW assembly technique[3] and mount them into a piston-cylinder pressure cell with electrical feedthroughs, capable of reaching temperatures below 1 K and magnetic fields above 18 T. The sample space in the pressure cell is filled with oil which results in a uniform transfer of pressure to the sample (see Methods and Extended Data Fig. 1 for details). We first examine the effect of applying pressure on a misaligned heterostructure for which no MSL effects are present, and find that the oil environment and application of pressure have virtually no effect on the electronic properties of the graphene. Figure 1b shows low-temperature transport acquired from a misaligned device (>2° relative alignment) at zero magnetic field both in vacuum (with no oil) and under high pressure (1 GPa). The high-density resistance grows by a small amount, less than 25 Ω, with pressure in this device. Other similar devices showed no measurable change or even slightly decreasing resistance under pressure, indicating that the field effect mobility of the encapsulated

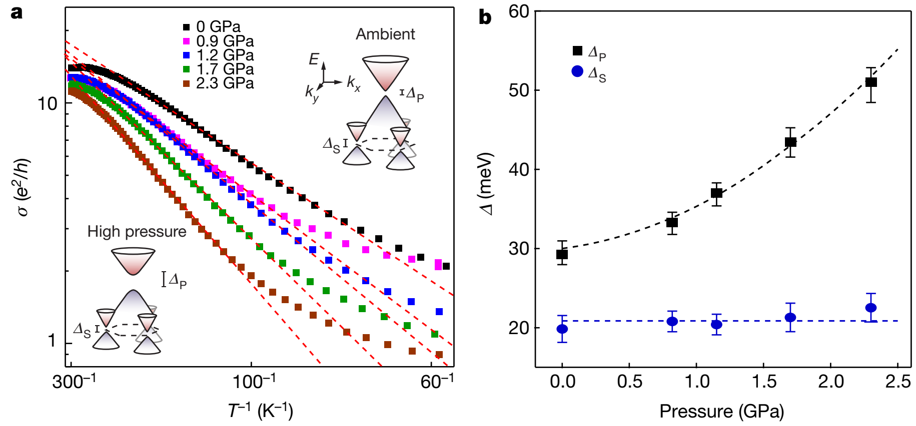


Fig. 3 | Bandgaps as a function of pressure. a, Arrhenius plot of the conductivity of the PDP as a function of inverse temperature for device P1 at various pressures. The slopes of the linear fits (red dotted lines) give the bandgap Δ at each pressure via $\sigma_{\mathrm{PDP}}(T) \propto \mathrm{e}^{-\Delta/(2kT)}$. Cartoon illustrations of the aligned graphene on BN band structure, including Δ_p and Δ_s and illustrating their dependence on pressure, are shown in the insets. Our model predicts SDPs at only one of the two valleys of the superlattice Brillouin zone. **b**, Bandgaps as a function of pressure. Δ_p grows with pressure, whereas Δ_s is relatively insensitive to pressure. Error bars in the gaps represent the uncertainty arising from determining the linear (thermally activated) regime for the fit (see Methods). The uncertainty in determining the pressure is smaller than the marker size. The dashed curves show the gaps predicted by the theoretical model using appropriately tuned deformations (see text).

device is largely insensitive to the application of pressure (see Extended Data Fig. 2). The PDP resistance is also not strongly modified, and there is no noticeable pressure dependence on the contact resistance (Fig. 1b inset).

Figure 1c shows similar transport measurements but acquired from an aligned MSL device. Under ambient pressure, the device shows excellent transport characteristics with moiré coupling evident by the appearance of two resistance peaks symmetrically located about the PDP at roughly $\pm 3.3 \times 10^{12}\,\mathrm{cm}^{-2}$, corresponding to a relative rotation angle of about 0.6°. Moreover, the large resistivities exceeding 25 kΩ at the PDP and SDP suggest sizeable bandgaps at those points in the band structure. Notably, as we apply high pressure up to 2.3 GPa, we find that the positions of the secondary peaks move symmetrically towards the PDP.

To understand this effect, we track the back gate capacitance per unit area $C_g = en/V_g = \varepsilon\varepsilon_0/t$, where e is the charge of the electron, ε is dielectric constant of BN, ε_0 is the vacuum permittivity, t is the thickness of the BN, V_g is the applied gate bias and n is the carrier density. We measure the density in two ways, both from the Hall effect and from magnetoresistance oscillations (Fig. 2a and b; see Methods). All devices exhibit a universal increase of C_g with pressure of roughly 9% per GPa, independent of their relative alignment (Fig. 2c), which must arise through a decrease in t and/or an increase in ε. To deconvolve the two, we have performed ab initio simulations of bulk BN multilayers under pressure, using the local density approximation (LDA) and find approximately 2.5% compression per GPa (green curve in the top inset of Fig. 2c), in good quantitative agreement with previous X-ray diffraction measurements[22] (reproduced in the red curve). The remaining increase in gate capacitance is therefore attributed to an increase in the dielectric constant of the BN of roughly 6% per GPa (blue curve in the bottom inset of Fig. 2c). Taken together, this suggests that pressure can sensitively tune both the interlayer spacing between layered 2D materials and their dielectric properties[23]. Returning to the transport measurements of the MSL device, we find that the three resistance peaks align exactly at all pressures when replotted against charge carrier density n (Fig. 1c inset), suggesting that although the graphene and BN layers move closer together, the relative rotation angle and moiré period remain fixed under pressure.

A second notable feature of the MSL transport is that the resistance at the PDP grows strongly with pressure, increasing by roughly 100 kΩ between 0 GPa and 2.3 GPa. We investigate the PDP response in more detail by measuring its temperature dependence. Figure 3a shows an Arrhenius plot of the PDP conductivity, σ_{PDP}, versus inverse temperature for various pressures, where for each pressure a linear fit to the thermally activated regime (red dashed lines) gives a measure of the activation gap (see Methods). The resulting PDP gap, Δ_p, is shown versus pressure in Fig. 3b (square markers). The gap is found to increase superlinearly with pressure, and is enhanced by nearly a factor of 2 at the highest pressure studied for the device presented here with relative rotation angle about 0.6°. We similarly measure the pressure dependence of the valence band SDP gap (Extended Data Fig. 3), Δ_s, plotted with filled circles in Fig. 3b. In contrast to the PDP gap, the SDP gap is nearly unresponsive to pressure. The insets of Fig. 3a schematically illustrate the inferred band-structure modifications with interlayer spacing, showing a growing Δ_p but a fixed Δ_s.

As a simple approximation, we may consider the increasing PDP gap with pressure to result from increasing MSL coupling owing to the decreasing interlayer spacing. In this case, we would expect that the SDP behaviour might respond in a similar way, and from this perspective its insensitivity to pressure is surprising. We note, however, that despite considerable effort[9,10,12,13,24–32], consensus is still lacking as to the exact origin of these bandgaps. Lattice-scale deformations (in-plane strains and out-of-plane corrugations) of the graphene layer are expected to play an important role[10,29–33], but the exact equilibrium structure of graphene in contact with BN remains poorly understood, including whether these deformations even exist in fully BN-encapsulated devices[10]. In an effort to understand the behaviour that we observe, we therefore first consider a rigid graphene lattice and examine the effects of pressure on the heterostructure by using a combination of ab initio and analytical models (see Methods and Supplementary Information). The interlayer electronic coupling between the graphene and BN, $V_0 = \tilde{V} e^{-\beta(z_0 - z_r)}$, is highly sensitive to pressure, with ab initio predictions indicating that the average interlayer spacing z_0 should decrease by about 0.07 Å GPa^{-1}. Here $\tilde{V} \approx 18$ meV is the interlayer electronic coupling, $z_r \approx 3.35$ Å is the equilibrium average interlayer spacing between the graphene and BN, and $\beta \approx 3.2$ Å^{-1} quantifies the rate of increase of interlayer coupling when the spacing is reduced. Generically, both Δ_p and Δ_s should scale proportionally to V_0 under applied pressure, but this is in stark contrast to our experimental observations in which Δ_s in particular exhibits little pressure dependence. Even at 0 GPa, a rigid model does not properly capture the experimentally observed hierarchy of the gaps, predicting a large Δ_s and a $\Delta_p \approx 0$ (Fig. 4a).

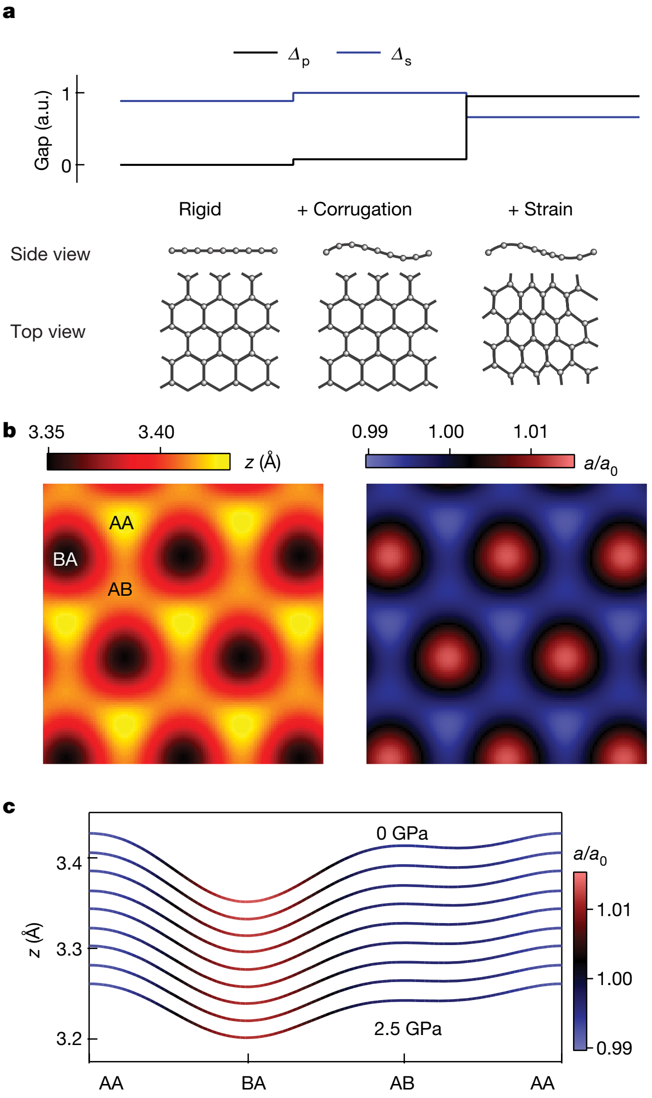


Fig. 4 | Modelling of the bandgaps. a, Relative contributions to Δ_p and Δ_s from lattice corrugations and strains. Strain is necessary to open a sizeable Δ_p, while deformations can substantially reduce Δ_s. Schematics of the graphene lattice structure for each case are illustrated below. **b**, Maps of the out-of-plane corrugations (left) and in-plane lattice strains (right) at 0 GPa used to model the gaps in Fig. 3b. **c**, Models for lattice corrugations through the high-symmetry points of the moiré as a function of pressure based on ab initio calculation inputs in BN-encapsulated graphene with one aligned interface. The left axis labels the atomic spacing between the BN and graphene lattices. The colour represents the magnitude of in-plane lattice strains needed to model the gaps in Fig. 3b.

Next we consider the potential role of atomic-scale graphene deformations. By considering realistic values of the graphene elastic deformation potential, our atomic structure models suggest that it is favourable for the graphene lattice both to corrugate out-of-plane and to be strained in-plane on the moiré scale to minimize the overall stacking potential with the BN substrate (Fig. 4b). These deformations break the sublattice symmetry of the graphene, resulting in a finite average mass term in the Hamiltonian which opens a sizeable Δ_p. There are also moiré-induced electrostatic potentials and pseudomagnetic fields that contribute to Δ_s in addition to the mass term. Including corrugations alone opens a small Δ_p at 0 GPa, but still does not recover the observed gap hierarchy of $\Delta_p > \Delta_s$ (Fig. 4a). By additionally including strains, we are able to recover the observed 0 GPa hierarchy of the gaps (Fig. 4a) as well as the superlinear increase in Δ_p and the flat Δ_s response with pressure. We find that we can quantitatively reproduce our experimental gap results with an analytical model for aligned graphene on BN which accounts for physically reasonable corrugations and strains consistent with predictions from ab initio modelling (dashed curves in Fig. 3b; see Methods and Supplementary Information for details of the model).

Although more experimental and theoretical work is necessary to understand the exact equilibrium structure of aligned graphene on BN, the varying evolution of the gaps with pressure directly rules out the possibility of a rigid graphene lattice in these structures, and moreover demonstrates that these gaps are of fundamentally different origin as the SDP cone is not a true replica of the PDP cone. This suggests the possibility of independent control over the magnitude of the two gaps, as well as other features of the moiré band structure and magnetoresponse, by selectively engineering specific lattice deformations. Additionally, sufficiently strong enhancement of the interlayer coupling could drive a phase transition to the fully commensurate (lattice-matched) stacking configuration[33], marked by the absence of a MSL but the emergence of strong sublattice-symmetry breaking in the graphene and a gap many times kT, where T is room temperature, at the Dirac point[26]. Even in the incommensurate configuration, extrapolating our results to 9 GPa (beyond which a phase transition in the crystal structure of the BN is expected) suggests that gaps as large as about 250 meV may be realized. This provides a potential new route to the long-pursued goal of engineering technologically relevant bandgaps in high-quality monolayer graphene. More generally, our results indicate that a wide variety of vdW heterostructure properties should be tunable by controlling the interlayer coupling strength with pressure, particularly those with emergent electronic properties driven by interlayer interactions. For example, sufficiently strong coupling to transition metal dichalcogenides with large spin–orbit coupling may drive topologically non-trivial gaps in graphene supporting a zero-field quantum spin Hall effect, and strong coupling to a vdW magnetic insulator may drive a quantum anomalous Hall effect. Even in simple vdW structures consisting of multilayers of the same material, fundamental features of the system such as the bandgap or the nature of properties such as superconductivity or magnetism scale with the interlayer coupling. Our demonstration of the tunability of interlayer interactions suggests that a wide variety of new device phenomena may now become experimentally accessible in 2D vdW structures.

Online content

Any Methods, including any statements of data availability and Nature Research reporting summaries, along with any additional references and Source Data files, are available in the online version of the paper at https://doi.org/10.1038/s41586-018-0107-1.

Received: 21 October 2017; Accepted: 14 February 2018;
Published online 16 May 2018.

1. Geim, A. K. & Grigorieva, I. V. Van der Waals heterostructures. *Nature* **499**, 419–425 (2013).
2. Novoselov, K. S., Mishchenko, A., Carvalho, A. & Castro Neto, A. H. 2D materials and van der Waals heterostructures. *Science* **353**, aac9439 (2016).
3. Wang, L. et al. One-dimensional electrical contact to a two-dimensional material. *Science* **342**, 614–617 (2013).
4. Mishchenko, A. et al. Twist-controlled resonant tunnelling in graphene/boron nitride/graphene heterostructures. *Nat. Nanotech.* **9**, 808–813 (2014).
5. Yu, H., Wang, Y., Tong, Q., Xu, X. & Yao, W. Anomalous light cones and valley optical selection rules of interlayer excitons in twisted heterobilayers. *Phys. Rev. Lett.* **115**, 187002 (2015).
6. Yankowitz, M. et al. Emergence of superlattice Dirac points in graphene on hexagonal boron nitride. *Nat. Phys.* **8**, 382–386 (2012).
7. Ponomarenko, L. A. et al. Cloning of Dirac fermions in graphene superlattices. *Nature* **497**, 594–597 (2013).
8. Dean, C. R. et al. Hofstadter's butterfly and the fractal quantum Hall effect in moiré superlattices. *Nature* **497**, 598–602 (2013).
9. Hunt, B. et al. Massive Dirac fermions and Hofstadter butterfly in a van der Waals heterostructure. *Science* **340**, 1427–1430 (2013).
10. Woods, C. R. et al. Commensurate–incommensurate transition in graphene on hexagonal boron nitride. *Nat. Phys.* **10**, 451–456 (2014).
11. Gorbachev, R. V. et al. Detecting topological currents in graphene superlattices. *Science* **346**, 448–451 (2014).
12. Wang, L. et al. Evidence for a fractional fractal quantum Hall effect in graphene superlattices. *Science* **350**, 1231–1234 (2015).
13. Wang, E. et al. Gaps induced by inversion symmetry breaking and second-generation Dirac cones in graphene/hexagonal boron nitride. *Nat. Phys.* **12**, 1111–1115 (2016).
14. Spanton, E. M. et al. Observation of fractional Chern insulators in a van der Waals heterostructure. *Science* **360**, 62–66 (2018).
15. Dean, C. R. et al. Boron nitride substrates for high-quality graphene electronics. *Nat. Nanotech.* **5**, 722–726 (2010).
16. Park, C.-H., Yang, L., Son, Y.-W., Cohen, M. L. & Louie, S. G. New generation of massless Dirac fermions in graphene under external periodic potentials. *Phys. Rev. Lett.* **101**, 126804 (2008).
17. Chen, G. et al. Emergence of tertiary Dirac points in graphene moiré superlattices. *Nano Lett.* **17**, 3576–3581 (2017).
18. Kim, K. et al. Van der Waals heterostructures with high accuracy rotational alignment. *Nano Lett.* **16**, 1989–1995 (2016).
19. Woods, C. R. et al. Macroscopic self-reorientation of interacting two-dimensional crystals. *Nat. Commun.* **7**, 10800 (2016).
20. Chari, T., Ribeiro-Palau, R., Dean, C. R. & Shepard, K. Resistivity of rotated graphite–graphene contacts. *Nano Lett.* **16**, 4477–4482 (2016).
21. Koren, E. et al. Coherent commensurate electronic states at the interface between misoriented graphene layers. *Nat. Nanotech.* **11**, 752–757 (2016).
22. Solozhenko, V. L., Will, G. & Elf, F. Isothermal compression of hexagonal grpahite-like boron nitride up to 12 GPa. *Solid State Commun.* **96**, 1–3 (1995).
23. Chen, Y. et al. Pressurizing field-effect transistors of few-layer MoS_2 in a diamond anvil cell. *Nano Lett.* **17**, 194–199 (2017).
24. Chen, Z.-G. et al. Observation of an intrinsic bandgap and Landau level renormalization in graphene/boron-nitride heterostructures. *Nat. Commun.* **5**, 4461 (2014).
25. Song, J. C. W., Shytov, A. V. & Levitov, L. S. Electron interactions and gap opening in graphene superlattices. *Phys. Rev. Lett.* **111**, 266801 (2013).
26. Bokdam, M., Amlaki, T., Brocks, G. & Kelly, P. J. Band gaps in incommensurable graphene on hexagonal boron nitride. *Phys. Rev. B* **89**, 201404(R) (2014).
27. Moon, P. & Koshino, M. Electronic properties of graphene/hexagonal-boron-nitride moiré superlattice. *Phys. Rev. B* **90**, 155406 (2014).
28. Wallbank, J. R., Mucha-Kruczynski, M., Chen, X. & Fal'ko, V. I. Moiré superlattice effects in graphene/boron-nitride van der Waals heterostructures. *Ann. Phys.* **527**, 359–376 (2015).
29. Jung, J., DaSilva, A. M., MacDonald, A. H. & Adam, S. Origin of band gaps in graphene on hexagonal boron nitride. *Nat. Commun.* **6**, 6308 (2015).
30. San-Jose, P., Gutiérrez-Rubio, A., Sturla, M. & Guinea, F. Spontaneous strains and gap in graphene on boron nitride. *Phys. Rev. B* **90**, 075428 (2014).
31. Slotman, G. et al. Effect of structural relaxation on the electronic structure of graphene on hexagonal boron nitride. *Phys. Rev. Lett.* **115**, 186801 (2015).
32. Jung, J. et al. Moiré band model and band gaps of graphene on hexagonal boron nitride. *Phys. Rev. B* **96**, 085442 (2017).
33. Yankowitz, M., Watanabe, K., Taniguchi, T., San-Jose, P. & LeRoy, B. J. Pressure-induced commensurate stacking of graphene on boron nitride. *Nat. Commun.* **7**, 13168 (2016).

Acknowledgements We thank P. San-Jose, J. Song, A. Shytov, L. Levitov, J. Wallbank and P. Moon for theoretical discussions. This work was supported by the National Science Foundation (NSF) (DMR-1462383). C.R.D. acknowledges partial support from the David and Lucille Packard foundation. Development of the device concept and fabrication process was partially supported by the NSF MRSEC program through Columbia in the Center for Precision Assembly of Superstratic and Superatomic Solids (DMR-1420634). We acknowledge S. Tozer for use of his 16 T PPMS which is partially supported as part of the Center for Actinide Science and Technology (CAST), an Energy Frontier Research Center (EFRC) funded by the Department of Energy, Office of Science, Basic Energy Sciences under award no. DE-SC0016568. A portion of this work was performed at the National High Magnetic Field Laboratory, which is supported by NSF Cooperative Agreement no. DMR-1157490, the State of Florida and the US Department of Energy and additionally provided support for pressure cell development through User Collaboration Grant Program (UCGP) funding. J.J. and N.L. were supported by the Korean NRF through grant NRF-2016R1A2B4010105 and Korean Research Fellowship grant NRF-2016H1D3A1023826, and B.L.C. was supported by grant NRF-2017R1D1A1B03035932. E.L. and S.A. are supported by the National Research Foundation of Singapore under its Fellowship program (NRF-NRFF2012-01) and the Singapore Ministry of Education AcRF Tier 2 (MOE2017-T2-2-140). K.W. and T.T. acknowledge support from the Elemental Strategy Initiative conducted by the MEXT, Japan, and JSPS KAKENHI grant no. JP15K21722.

Author Contributions M.Y. and C.R.D. conceived the experiment. M.Y. fabricated the samples and analysed the data. M.Y. and D.G. performed the experiments. J.J., E.L. and S.A. developed the theory. N.L. and B.L.C. calculated the ab initio potentials. K.W. and T.T. grew the hBN crystals. C.R.D. advised on the experiments. All authors participated in writing the paper.

Competing interests The authors declare no competing interests.

Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0107-1.
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0107-1.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to C.R.D.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

Device structure and application of pressure. All devices in this study consist of monolayer graphene encapsulated between two layers of BN. The BN layers were 20–60 nm thick, although our results did not depend on this in any important way. The encapsulated stack sits on a flake of graphite that acts as a local back gate, and the entire stack rests on a Si/SiO$_2$ wafer (Fig. 1a). Extended Data Fig. 1a shows an image of a completed heterostructure on a transfer slide. For aligned samples, the graphene (outlined with a dashed white line) was either intentionally aligned optically to one of the BN layers using crystalline edges of the flakes[7] or incidentally through self-rotation to an aligned position, which has lowest stacking potential energy, upon heating during the transfer process[12,19]. The final device was partially etched into a Hall bar geometry, leaving some of the bottom BN unetched to prevent the metal contacts from shorting to the graphite gate (Extended Data Fig. 1b). We make one-dimensional electrical contact using standard reactive ion etching and electron-beam patterning and deposition techniques[3]. The Hall bar was intentionally kept small (approximately 6 μm by 2 μm) to keep the pressure as uniform as possible across the entire device. The Si/SiO$_2$ substrate is diced to approximately 2 mm by 2 mm to fit into the inner bore of the pressure cell (Extended Data Fig. 1c).

To prepare the pressure cell, a clean metal stage (Extended Data Fig. 1d) is threaded with insulated copper 75 μm wires with tinned ends, which are epoxied in place using Stycast 2850 FT using 24LV catalyst (Extended Data Fig. 1e). A ruby crystal is glued to the tip of a thin optical fibre that is fixed in place by the wires for in situ pressure calibration. The sample is then glued above the fibre (Extended Data Fig. 1f). Flexible 15 μm Pt wires are soldered to the copper wires and affixed by hand to the gold sample contacts using Dupont 4929 N silver paste (Extended Data Fig. 1g). Next, a Teflon cup is filled with the pressure medium (Daphne 7373 or 7474 oil[34,35]) and carefully fitted over the sample and onto the stage (Extended Data Fig. 1h). The sample is then completely encapsulated in oil. For the special case of vacuum measurements (0 GPa), no oil is loaded into the Teflon cup. The stage/Teflon cup is then fitted into the inner bore of a piston cylinder cell (Extended Data Fig. 1i) and a hydraulic press is used to compress the top of the Teflon cup (Extended Data Fig. 1j), which is held in place by a locking nut. The cell is then affixed to the end of a probe (Extended Data Fig. 1k) for low-temperature, high-field measurements. The pressure was determined by measuring the fluorescence response of the ruby crystal at both room temperature and low temperature[36,37]. Changing the pressure requires warming the cell to room temperature, adding or removing load with the hydraulic press, and re-cooling the cell. Although the pressure in the cell is uniform, the Young's modulus in the the out-of-plane stacking direction (*c* axis) of vdW crystals is typically a few orders of magnitude smaller than in-plane[38]; hence the pressure primarily results in a *c*-axis compression.

Although the oil does not noticeably influence the electronic properties of the device, special care must be taken to account for its presence during measurements. The oil freezes at around 200 K as it is cooled at ambient pressure, and the freezing point moves to higher temperatures at higher starting loads. The pressure in the cell also drops as the oil cools inside the cell, with a larger relative drop in pressure at smaller initial loads. For example, a starting load below about 0.3 GPa at room temperature will result in nearly ambient conditions at low temperature, whereas at pressures above 2 GPa there is virtually no change between the room-temperature and low-temperature pressures. We have found that the primary consequence of this effect is that starting at room-temperature loads of roughly 0.5 GPa or less is dangerous for the device, as the stack is typically torn when the oil freezes below these pressures. However, above these pressures the devices always survive the cooling, and we have not noticed any effect of the oil freezing in transport measurements. Finally, special care is taken to account for the large thermal load of the pressure cell when performing temperature sweeps to measure bandgaps. The temperature is swept slowly to keep the sample as close to equilibrium as possible, ideally at 0.5 K min^{-1}, and no faster than 1.5 K min^{-1}.

Apart from the change in PDP resistance, we typically do not observe much modification of the electronic response of our devices with pressure. For device P2 (results shown in Fig. 1b), an increase of 15–25 Ω is observed at high density under pressure, although this is not always the case. For example, Extended Data Fig. 2 shows similar transport measurements for device P3, in which we observe virtually no change in the device resistivity for hole carriers, and a decrease in the resistivity for electron carriers. In any case, the change in high-density response is always very small compared with the effect of pressure at the PDP.

Extraction of bandgaps and capacitance. We extract the bandgaps Δ of the PDP and SDP at each pressure according to the thermally activated response where $\sigma_{\mathrm{PDP}}(T) \propto \mathrm{e}^{-\Delta/(2kT)}$, where k is the Boltzmann constant. The error bars on the gap values in Fig. 3b reflect the uncertainty in determining the temperature range over which the Arrhenius plot shows a linear response. The upper and lower error boundaries correspond to the magnitude of the gap sizes extracted when fitting a line to the minimum and maximum temperature ranges over which the dispersion in Fig. 3a could be reasonably judged as linear. The error arising from the uncertainty determining the extraction range is much larger than the standard error in any individual linear regression. Figure 3a shows this fit for the PDP in device P1 as a function of pressure, and Extended Data Fig. 3 similarly shows this fit for the valence-band SDP, demonstrating that the latter gap does not change appreciably with pressure.

To find the gate capacitance, we extract $n(V_g)$ for each pressure through either the dispersion of the quantum Hall states in high magnetic field as $n = \nu eB/h$, where ν is the filling factor (Fig. 2a), or from the low-field Hall resistance R_{xy} before the onset of strong Shubnikov–de Haas oscillations as $n = B/(eR_{xy})$ (Fig. 2b). Quantum Hall states move symmetrically closer to the PDP with increasing pressure, and the slope of the Hall resistance decreases with pressure, both implying a growing $n(V_g)$ with increasing pressure. In Fig. 2c, C_g is normalized to its measured value at 0 GPa to account for the different thicknesses of the bottom BN layers across the different devices. Additionally, the ab initio value of ε at 0 GPa is normalized to match the average experimental value of about 3. Our ab initio simulations also confirm that the BN dielectric constant should grow with pressure, although the effect is predicted to be a few times smaller (green curve in bottom inset), probably owing to the absence of many-body and excitonic effects in our calculations (see Supplementary Information for further discussion).

Gap evolution in other devices. We have measured a total of six devices as a function of pressure. Extended Data Fig. 4a plots the gap of the PDP, where the data from device P1 are copied from the main text. Error bars are left off for clarity, but are similar in magnitude to those in Fig. 3b. All devices except P1 are misaligned more than about 2° from both BNs, as we do not observe any SDPs within the experimentally accessible carrier density ranges (therefore the exact rotation angle is unknown). Previous work has demonstrated that devices with misalignment angles as large as about 5° can exhibit bandgaps in transport[9]. In that study, the misalignment angle was determined by scanning tunnelling topography measurements, but this is not possible in our devices owing to the presence of the encapsulating top BN layer. Nevertheless, some of these devices (P3 and P4) show clear activation gaps, with insulating behaviour at low temperatures and a decrease of the device conductivity by nearly an order of magnitude in the thermally activated regime at high temperatures (Extended Data Fig. 5c, d). Two of the devices (P2 and P5) only exhibit activated behaviour over a much smaller range, but we can nevertheless perform the gap fit over the regime that is activated, with the understanding that this number may not be an accurate measure of the bandgap. These gaps are denoted with triangle markers on the plot in Extended Data Fig. 4a. Strikingly, in every device studied the gap grows with increasing pressure, and the magnitude of the gap enhancement seems to scale roughly with the magnitude of the gap at 0 GPa. This is consistent with our proposed model in which the gap depends on the interlayer electronic coupling between the graphene and BN, as well as lattice deformations resulting from structural relaxation of the graphene to minimize the stacking energy, the latter of which should diminish quickly with increasing misalignment angle. Although we expect that the misalignment angle in these devices is larger for devices with smaller gaps at 0 GPa, this remains an open question owing to our inability to directly probe the angular misalignments with each BN.

We also measured a second aligned sample (device P6), with an estimated misalignment angle of about 0.1°. This device was considerably more disordered than all the other devices in this study, and as a result the onset of the variable-range hopping regime was at higher temperature. Consequentially, the thermally activated regime spanned only a small range of temperature, and therefore the gap extraction is more ambiguous than for the other aligned device (P1). Nevertheless, the Δ_{p} showed a clear enhancement with pressure, whereas Δ_{s} seemed relatively insensitive to pressure (Extended Data Fig. 4b). This behaviour is consistent with the clean aligned device (P1) presented in the main text, providing further evidence of the fundamental difference between the primary and secondary Dirac points.

The enhancement of the bandgap does not depend on the history of the pressure to which the device has been exposed, but rather responds directly to the pressure the device is under at the time of measurement. To illustrate this, Extended Data Fig. 4a annotates the order in which the gaps were acquired. In no case does the measured gap fall out of the anticipated sequence. We have also examined the PDP conductivity at the lowest temperatures measured, but we find that this is not a good measure of the bandgap as it is additionally influenced by the effects of residual disorder in the variable-range hopping regime[9]. Extended Data Fig. 5 tracks the PDP conductivity across all measured devices below 2 K. Typically the resistance of the PDP is higher at low temperature for higher pressures, however this is not always the case. For devices P1 and P3, the low-temperature conductivity becomes out of sequence upon unloading pressure, or in reloading pressure after previously pressure cycling and returning to 0 GPa (see the annotations in Extended Data Fig. 4a for the order of curve acquisition). Even in device P4, for which the pressure was loaded up uniformly, the low-temperature conductivity still depends non-monotonically on pressure. This suggests that small changes in the magnitude of disorder in the device and potentially even the details of its microscopic organization can ultimately influence how insulating the device becomes. However, no matter how the PDP conductivity behaves at low temperature, the bandgap is always enhanced by pressure, suggesting that the gap enhancement is a robust property of the heterostructure and depends most critically on the interlayer interaction strength between the graphene and the BN rather than small changes in device disorder.

Effects of pressure in the quantum Hall regime. The disorder in these devices can also be characterized by considering the density n necessary to switch the Hall resistance R_{xy} from positive to negative (Extended Data Fig. 6a), as this gives a measure of the effective magnitude of the electron–hole puddles in the bulk. Extended Data Fig. 6b shows that the device disorder is not strongly dependent on pressure, growing by less than the extraction uncertainty over the applied pressure range. Curiously, pressure serves to qualitatively improve the quantum Hall response at high magnetic field. Extended Data Fig. 6c shows one such example, where both integer and fractional quantum Hall states develop more clearly under pressure. This effect is observed in every device examined. Surprisingly, we find that this improvement persists even after the pressure is released (Extended Data Fig. 6d), suggesting that this improvement is an irreversible effect. At present, we may only speculate as to the nature of this effect.

First, it is important to note that despite the high electronic quality of the devices, only the main sequence of IQH gaps, at $\nu = -2, -6, -10$ and so on, are typically fully developed in initial measurements in vacuum, with the symmetry-broken gaps still developing (Extended Data Fig. 6d). The poor development of the QHE despite the high mobility is symptomatic of the device geometry used here (Extended Data Fig. 1b), where the metal contacts sit above the graphite gate. Although the reason for this is currently not well understood, the situation may be analogous to previous observations in GaAs, in which a partial reduction of the carrier concentration in the graphene just in front of the contacts can lead to a depletion region along the boundary, which may impede the ability to observe well-developed quantum Hall plateaus[39]. This problem has been addressed previously by leaving Si-gated regions of the graphene to act as electrical contacts to the Hall bar, where the density at the boundary can vary more smoothly[40]. However, the devices in this study were intentionally kept small to ensure the pressure was as uniform as possible across the device, and so this geometry was not used.

Applying pressure may, for instance, provide a way to reduce this depletion region at the contacts, permitting better coupling to the quantum Hall edge modes. This could arise because of self-cleaning of contaminants along the contact boundaries, or if the metal making edge contact to the graphene forms a better bond to the graphene edge and lowers the work function mismatch at the boundary. Both of these effects could in principle persist even after the pressure is released. We find that the contact resistance is not noticeably modified by pressure at $B = 0$ T (Fig. 1b), pointing to an improvement arising from the detailed electrostatics at the contact barrier in high magnetic field, or to a reduction of non-local contamination which may be most relevant in the quantum Hall regime. Further study is necessary to understand the exact nature of this effect.

Finally, we examine the effects of pressure on the Hofstadter butterfly spectrum of the aligned device P1. Extended Data Fig. 7 shows the high-field response at ambient conditions (Extended Data Fig. 7a) and at 2.3 GPa (Extended Data Fig. 7b). As pressure enhances the coupling to the MSL, it should consequentially modify the relative strength of magnetotransport features arising from the secondary Dirac cones. We observe that under pressure the longitudinal magnetoresistance becomes larger for features arising from the valence band SDP, in qualitative agreement with the expectation of a stronger MSL. However, this must be deconvolved with the overall change in the quantum Hall behaviour with pressure, and a full investigation of the dependence of integer and fractional quantum Hall gaps on pressure is outside the scope of this work.

Theoretical model. Graphene and BN are two-dimensional materials with hexagonal crystal structures, but with different lattice constants. When graphene is deposited on a BN substrate with near-perfect alignment (small twist angles), the slight lattice mismatch ($\varepsilon \approx 1.7\%$) leads to the formation of a moiré structure which is characterized by a wavelength that is one to two orders of magnitude larger than the graphene unit cell. As a consequence, the dynamics of the low-energy Dirac electrons of graphene are governed by the corresponding long-wavelength component of the electronic couplings. Similarly, the interlayer potentials and deformations in graphene/BN are also largely characterized by this long-wavelength theory. The dominant contributions are captured by the first harmonics of the moiré structure[28,29,41] characterized by a set of vectors $\boldsymbol{G}_m$ that are related to the original graphene lattice vectors $\boldsymbol{g}_m$ by

$$\boldsymbol{g}_m = \hat{\boldsymbol{R}}_{2\pi(m-1)/6}(0, g), \quad \boldsymbol{G}_m = [(1+\varepsilon) - \hat{\boldsymbol{R}}_\theta]\boldsymbol{g}_m \approx \varepsilon \boldsymbol{g}_m - \theta \hat{\boldsymbol{z}} \times \boldsymbol{g}_m \qquad (1)$$

where $m \in \{1, 2, \ldots, 6\}$, $\hat{\boldsymbol{R}}_\theta$ denotes a rotation by θ, and $g = 4\pi/(3a)$ is the length of graphene lattice vectors with $a \approx 1.42$ Å the carbon–carbon distance. The approximate sign indicates the approximation within the small twist angle limit ($\theta \ll 1$). We show in Supplementary Information that $\boldsymbol{G}_m$ defines the moiré Brillouin zone, whose lateral dimension is scaled by $\tilde{\varepsilon} = \sqrt{\varepsilon^2 + \theta^2}$ with respect to the graphene Brillouin zone and is less than about 5% for twist angles θ of less than about 2°. In the graphene/BN system which possesses triangular symmetry, we can define two periodic functions within the first harmonics: $f_1(\boldsymbol{r}) = \sum_m \exp(i\boldsymbol{G}_m \cdot \boldsymbol{r})$ which satisfies inversion and hexagonal symmetries, and $f_2(\boldsymbol{r}) = -i\sum_m (-1)^m \exp(i\boldsymbol{G}_m \cdot \boldsymbol{r})$ which is asymmetric under inversion.

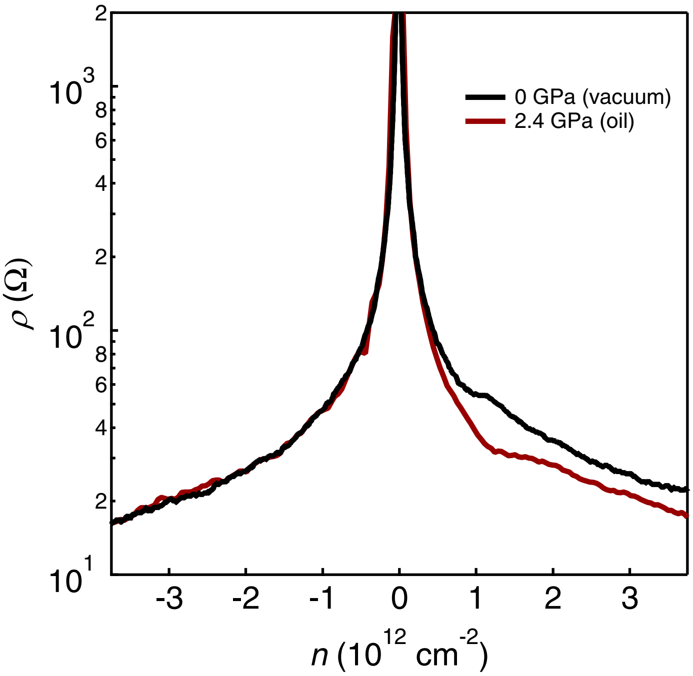
The low-energy Hamiltonian for the coupled graphene on BN system can be conveniently decomposed into a basis of $f_1(\boldsymbol{r})$ and $f_2(\boldsymbol{r})$. Consistent with previous results[32], we choose a reference frame in which the inversion-asymmetric couplings are an order of magnitude larger than the inversion-symmetric electronic couplings, which allows us to simplify the theoretical analysis considerably and obtain closed-form expressions for the primary and secondary gaps. Here we provide a qualitative discussion of the primary and secondary gaps, leaving the detailed analysis for Supplementary Information.

Both gaps have an exponential prefactor $e^{-\beta(z_0 - z_r)}$, which results in an increase in the gap arising from the increased coupling between the graphene and BN when the two layers are driven closer together under pressure. For Δ_p, we find that this exponential prefactor is modified by two terms. The first depends only on the corrugations and to leading order is given by the symmetric component of the out-of-plane deformations. The second term, proportional to the elastic deformability of the graphene, depends on both in-plane and out-of-plane deformations, but for realistic parameters depends only on the symmetric component of the deformations. As a result, to leading order Δ_p is determined by the symmetric in-plane and out-of-plane deformations and vanishes in the case of rigid graphene. In contrast, Δ_s is finite even for the rigid case without any deformations, and has additional terms arising from the electrostatic potentials and pseudomagnetic fields of the moiré structure which are not proportional to the exponential prefactor and provide a negative contribution to the magnitude of the gap. The experimental observation that $\Delta_p > \Delta_s$ implies that these negative terms are of comparable magnitude to the terms that scale exponentially with interlayer separation. Furthermore, Δ_s is controlled by both the symmetric and asymmetric components of the deformation, and to remain constant under pressure the difference between the symmetric and asymmetric deformations must increase with pressure. Additional analysis is provided in Supplementary Information.

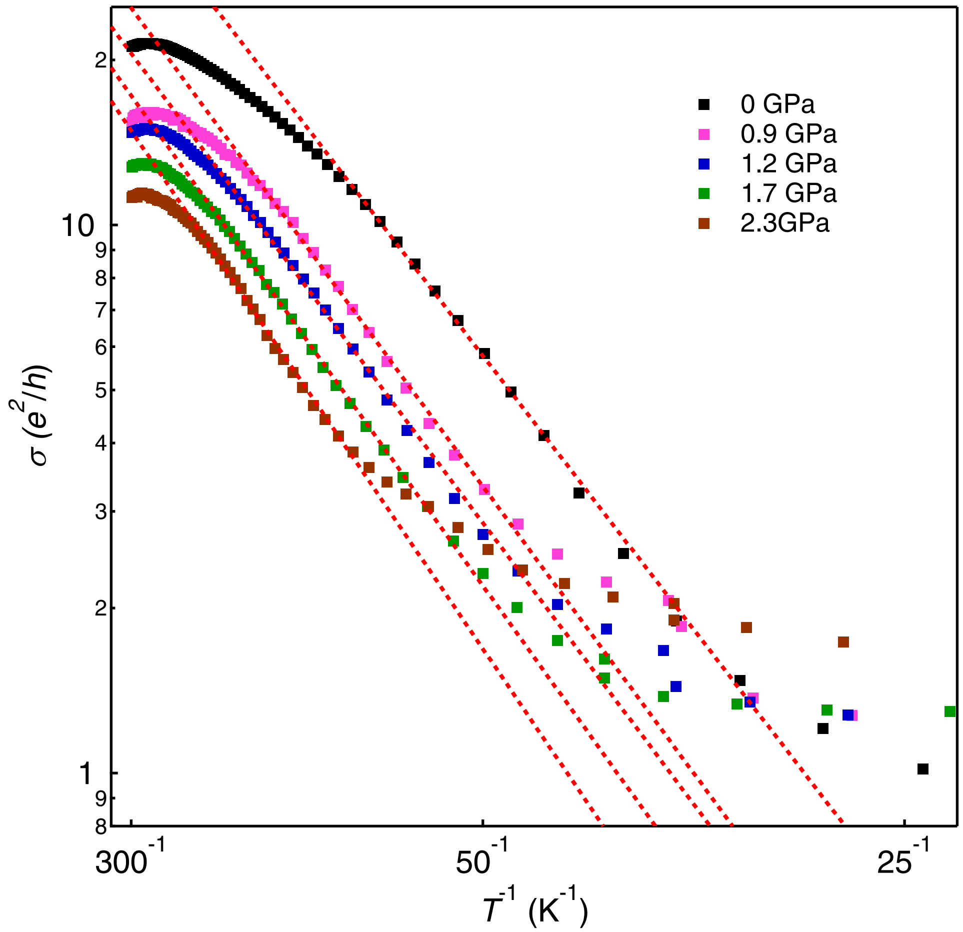
Data availability. The data presented in the figures and that support the other findings of this study are available from the corresponding author on reasonable request.

34. Murata, K., Yoshina, H., Yadav, H. O., Honda, Y. & Shirakawa, N. Pt resistor thermometry and pressure calibration in a clamped pressure cell with the medium, daphne 7373. *Rev. Sci. Instrum.* **68**, 2490–2493 (1997).
35. Murata, K. et al. Pressure transmitting medium Daphne 7474 solidifying at 3.7 gpa at room temperature. *Rev. Sci. Instrum.* **79**, 085101 (2008).
36. Piermarini, G. J., Block, S. & Barnett, J. D. Hydrostatic limits in liquids and solids to 100 kbar. *J. Appl. Phys.* **44**, 5377–5382 (1973).
37. Ragan, D. D., Gustavsen, R. & Schiferl, D. Calibration of the ruby R_1 and R_2 fluorescence shifts as a function of temperature from 0 to 600 K. *J. Appl. Phys.* **72**, 5539–5544 (1992).
38. Blakslee, O. L., Proctor, D. G., Seldin, E. J., Spence, G. B. & Weng, T. Elastic constants of compression-annealed pyrolytic grpahite. *J. Appl. Phys.* **41**, 3373–3382 (1970).
39. Weis, J. & von Klitzing, K. Metrology and microscopic picture of the integer quantum Hall effect. *Philos. Trans. R. Soc. Lond. A* **369**, 3954–3974 (2011).
40. Maher, P. et al. Tunable fractional quantum hall phases in bilayer graphene. *Science* **345**, 61–64 (2014).
41. Bistritzer, R. & MacDonald, A. H. Moiré bands in twisted double-layer graphene. *Proc. Natl Acad. Sci. USA* **108**, 12233–12237 (2011).

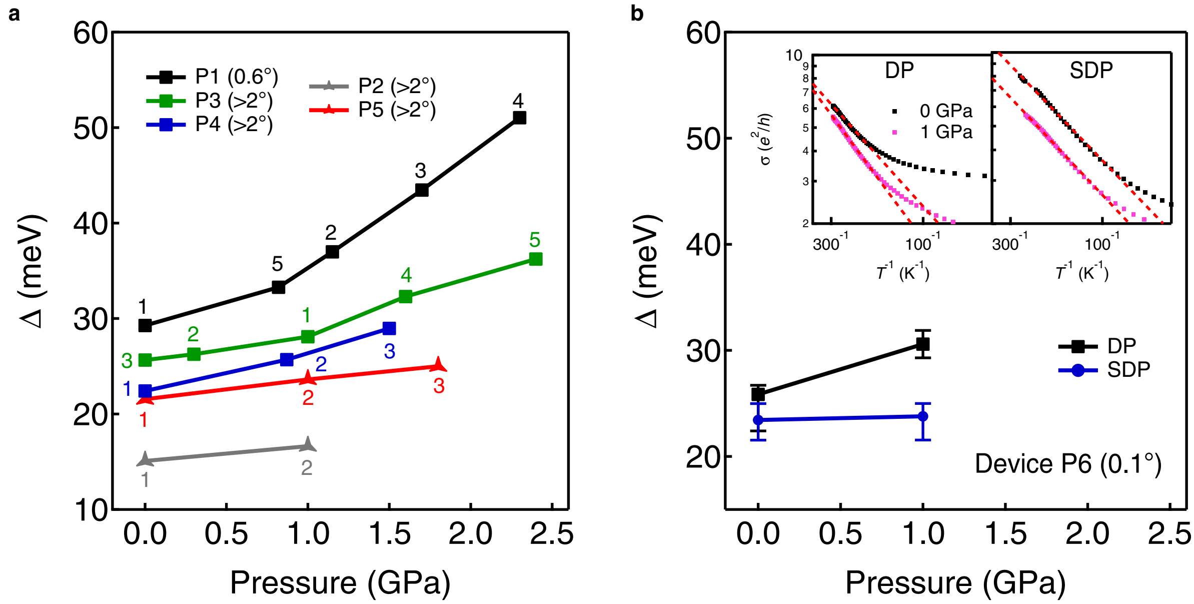
Extended Data Fig. 1 | Pictures of the set-up for pressure experiments. **a**, Graphene (outlined in white dashed line) is encapsulated between BN and placed on a graphite back gate. **b**, The device is shaped into a Hall bar and contacted for electrical measurements. **c**, The Si wafer is diced to approximate dimensions of 2 mm × 2 mm. **d**, A new pressure cell stage. **e**, The stage is threaded with wires which are epoxied in place. **f**, The sample is glued to the top of the stage. **g**, Wires are connected to the device electrodes by hand using silver paste. **h**, A Teflon cup is filled with oil and fitted over the sample and onto the stage. **i**, The sample is placed into a pressure cell. **j**, Pressure is loaded using a hydraulic press. **k**, The loaded pressure cell is attached to a probe for low-temperature magnetotransport measurements. See Methods for complete experimental details. The scale bars for **a** and **b** are 25 μm, and the scale bar for **c** is 2 mm.



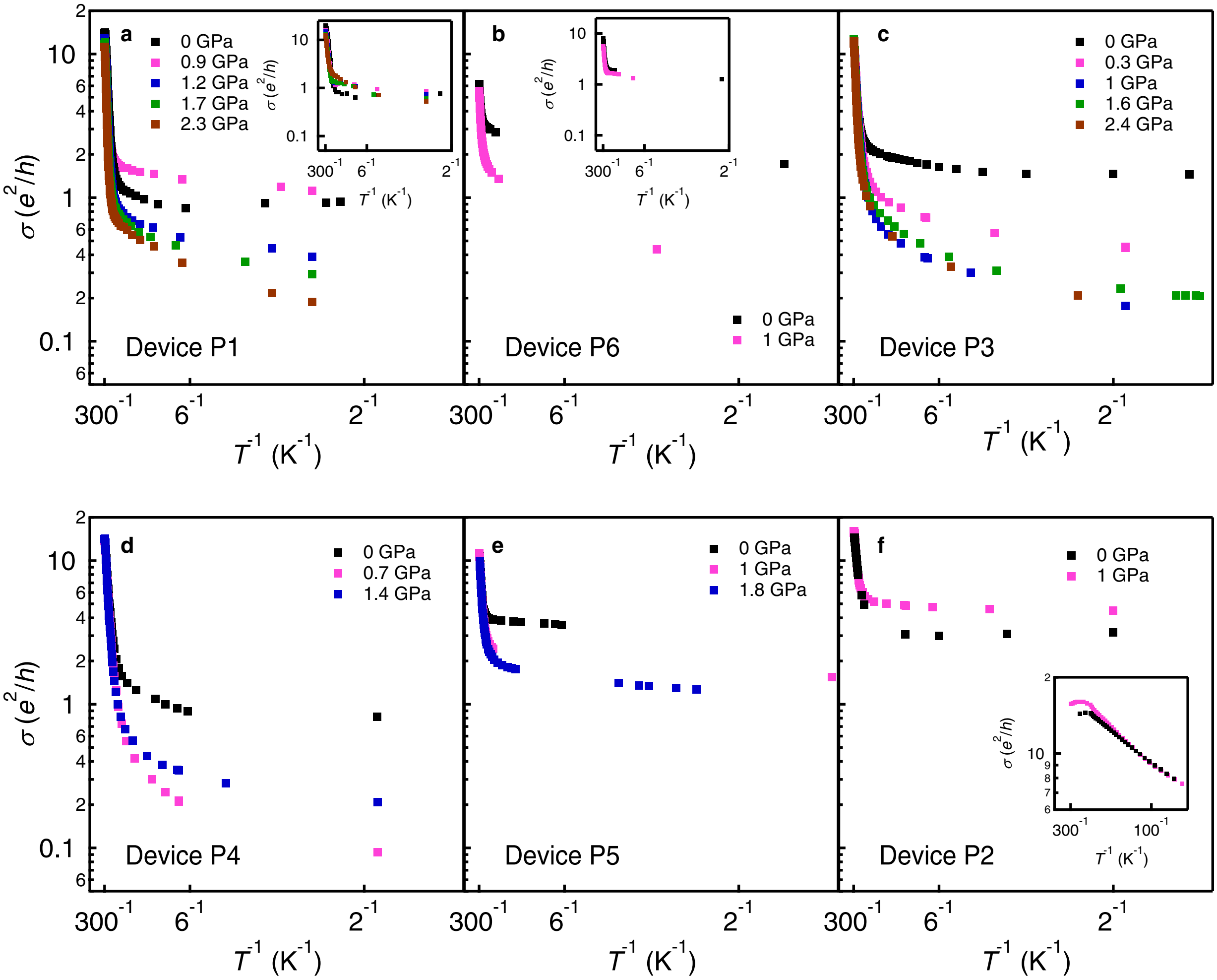
Extended Data Fig. 2 | Resistivity of device P3 at $B = 0$ T and $T = 2$ K. For the application of high pressure in this device, we observe virtually no change in the device resistivity for hole carriers, and a decrease in the resistivity for electron carriers.



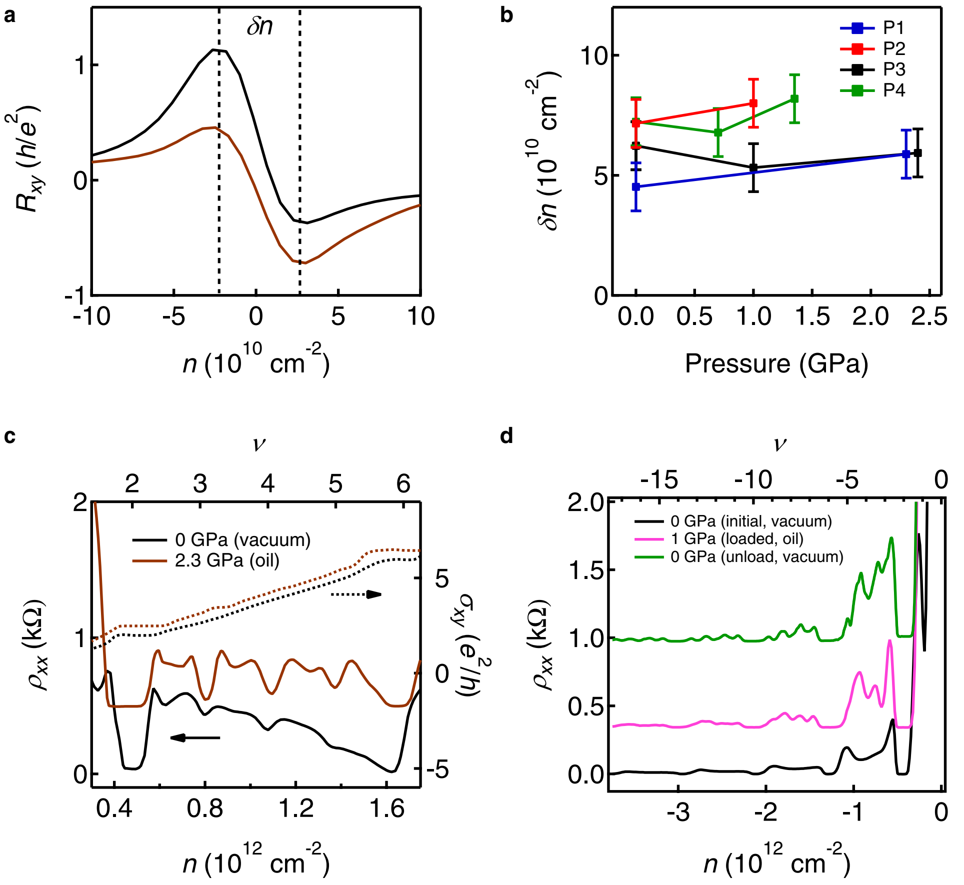
Extended Data Fig. 3 | Arrhenius plot for the SDP in device P1. The gap exhibits virtually no dependence on pressure.



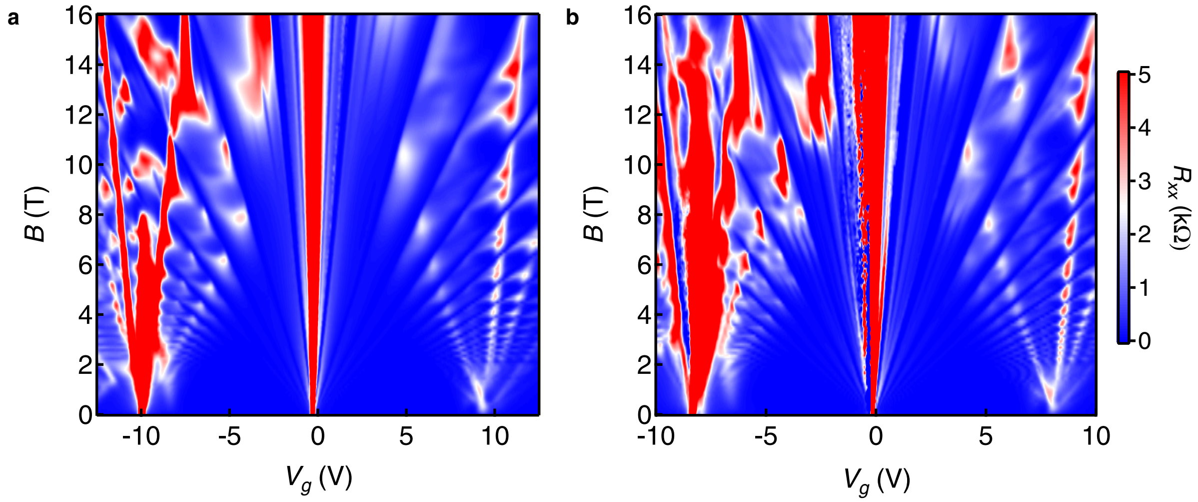
Extended Data Fig. 4 | Bandgaps as a function of pressure for all devices studied. a, Devices with square markers exhibit thermally activated behaviour over roughly an order of magnitude change in resistance. Devices with triangle markers exhibit thermally activated behaviour over a smaller range. The numbered labels represent the order in which the gaps were acquired. The gap magnitude depends only on the pressure the device is under at the time of measurement, and not on the history of the pressure that has been previously applied. Error bars in the gap fitting arising from the uncertainty identifying the thermally activated regime have been omitted for clarity, but are all less than 2.5 meV. **b**, Gap dependence in a second aligned device (P6), with misalignment of about 0.1°. This device also exhibits a growing Δ_p and constant Δ_s, but is much more disordered than device P1. Inset: Arrhenius plots for the PDP (left) and SDP (right) at 0 GPa and 1 GPa.



Extended Data Fig. 5 | Conductivity of the PDP over the full temperature range at various pressures for all devices studied. **a**, Device P1 is the aligned device discussed in the main text. Inset, conductivity of the SDP. **b**, Device P6 is a second aligned device, but is more highly disordered and therefore less resistive at low temperature. Inset, conductivity of the SDP. **c**–**f**, The remaining devices do not exhibit SDPs, and therefore the alignment to the BN layers is unknown. However, the devices in **c** and **d** still exhibit thermally activated behaviour over roughly an order of magnitude change in resistance, suggesting they are nearly aligned to a BN. The devices in **e** and **f** have a much smaller thermally activated regime. Pressure tends to have a smaller effect on the PDP conductivity for devices that are less resistive at 0 GPa. The inset in **f** shows the high-temperature response, demonstrating the weak effect of pressure in the thermally activated regime for the least resistive device. In **a**–**f**, although the PDP generally grows more insulating with higher pressure at low temperatures, this is not universally true, suggesting that the details of the low-temperature resistance depends more critically on the exact nature of disorder in the device. Devices are ordered by decreasing gap size as measured in the thermally activated regime.



Extended Data Fig. 6 | Disorder and the quantum Hall effect. **a**, Device disorder δn is measured as the n necessary to dope between the maximum and minimum R_{xy}, averaged over $B = 0.25$ T to 0.75 T. **b**, Density fluctuations as a function of pressure across different devices. The error bars denote the uncertainty in picking the peak positions of R_{xy}. **c**, ρ_{xx} (solid lines, left axis) and σ_{xy} (dotted lines, right axis) of device P1 at $B = 12.5$ T. Symmetry-broken integer quantum Hall states become much more clearly developed with pressure, and fractional quantum Hall states begin to emerge as well. Both are offset for clarity. **d**, ρ_{xx} of device P2 at $B = 9$ T, similarly showing an improvement in the quantum Hall response. Surprisingly, the improvement persists even after the pressure is released back to vacuum and the device is cleaned in solvents (green curve). Curves are vertically offset for clarity.



Extended Data Fig. 7 | Hofstadter butterfly as a function of pressure in device P1. Magnetotransport data acquired at $T = 2$ K and a pressure of: **a**, 0 GPa; **b**, 2.3 GPa.

LETTER

OPEN

https://doi.org/10.1038/s41586-018-0108-0

Genome sequence of the progenitor of wheat A subgenome *Triticum urartu*

Hong-Qing Ling[1,2,6]*, Bin Ma[3,6], Xiaoli Shi[1,6], Hui Liu[3,6], Lingli Dong[1,6], Hua Sun[1,6], Yinghao Cao[3], Qiang Gao[3], Shusong Zheng[1], Ye Li[1], Ying Yu[3], Huilong Du[2,3], Ming Qi[3], Yan Li[3], Hongwei Lu[2,3], Hua Yu[3], Yan Cui[1], Ning Wang[1], Chunlin Chen[1], Huilan Wu[1], Yan Zhao[1], Juncheng Zhang[1], Yiwen Li[1], Wenjuan Zhou[1], Bairu Zhang[1], Weijuan Hu[1], Michiel J. T. van Eijk[4], Jifeng Tang[4], Hanneke M. A. Witsenboer[4], Shancen Zhao[5], Zhensheng Li[1], Aimin Zhang[1]*, Daowen Wang[1,2]* & Chengzhi Liang[2,3]*

***Triticum urartu* (diploid, AA) is the progenitor of the A subgenome of tetraploid (*Triticum turgidum*, AABB) and hexaploid (*Triticum aestivum*, AABBDD) wheat[1,2]. Genomic studies of *T. urartu* have been useful for investigating the structure, function and evolution of polyploid wheat genomes. Here we report the generation of a high-quality genome sequence of *T. urartu* by combining bacterial artificial chromosome (BAC)-by-BAC sequencing, single molecule real-time whole-genome shotgun sequencing[3], linked reads and optical mapping[4,5]. We assembled seven chromosome-scale pseudomolecules and identified protein-coding genes, and we suggest a model for the evolution of *T. urartu* chromosomes. Comparative analyses with genomes of other grasses showed gene loss and amplification in the numbers of transposable elements in the *T. urartu* genome. Population genomics analysis of 147 *T. urartu* accessions from across the Fertile Crescent showed clustering of three groups, with differences in altitude and biostress, such as powdery mildew disease. The *T. urartu* genome assembly provides a valuable resource for studying genetic variation in wheat and related grasses, and promises to facilitate the discovery of genes that could be useful for wheat improvement.**

The genome of *T. urartu* (Tu) accession G1812 (PI428198) was sequenced and assembled (Extended Data Fig. 1a–c). The assembled contig sequences were 4.79 Gb with an N50 (the length N for which 50% of all bases in the sequences are in a sequence of length $L < \mathrm{N}$) of 344 kb and scaffold sequences were 4.86 Gb with an N50 of 3.67 Mb (Table 1), very close to the estimated genome size of 4.94 Gb[6]. We anchored 4.67 Gb (95.9%) of the scaffold sequences onto Tu chromosomes with a high density single nucleotide polymorphism (SNP) genetic map (Extended Data Fig. 1d), generating seven DNA pseudomolecules (Supplementary Data 1, Extended Data Fig. 1e). The high quality of the assembled sequences was confirmed at both the nucleotide and chromosome levels by comparison with previously published BAC sequences and the draft genome sequence of *Triticum aestivum* (Ta)[7] (Extended Data Fig. 2, Supplementary Data 2).

We predicted 41,507 protein-coding genes, including 37,516 high-confidence and 3,991 low-confidence genes (Extended Data Table 1a) using the Gramene pipeline[8]. On average, the genes have transcript length of 1,453 bp, protein length of 332 amino acids, and 4.5 exons per transcript, which were comparable to genes in other grasses[9,10] (Extended Data Table 1b). Approximately 88.18% of the predicted genes were assigned functional annotations (Extended Data Table 1c). We also predicted that 10,514 genes could produce alternatively spliced transcripts, with an average of 2.95 transcripts per gene. Moreover, we identified 31,269 microRNAs (miRNAs), 5,810 long non-coding RNAs (lncRNAs), 3,620 transfer RNAs (tRNAs), 80 ribosomal RNAs (rRNAs) and 2,519 small nuclear RNAs (snRNAs) throughout the genome (Extended Data Table 1d).

A total of 3.90 Gb (81.42%) of genome sequences was identified as repetitive elements, including 3.44 Gb (71.83%) of retrotransposons and 355 Mb (7.41%) of DNA transposons (Extended Data Table 1e). Among long-terminal repeat (LTR) retrotransposons, the Gypsy and Copia superfamilies comprised 42.71% and 24.30% of the genome, respectively. Further, we identified 48,370 intact Gypsy retrotransposons for which the peak of amplification bursts appeared at more than one million years ago (Ma) and 35,559 intact Copia retrotransposons with a peak less than 1 Ma (Fig. 1a). We also identified 121,792 solo-LTR/Gypsy and 44,349 solo-LTR/Copia elements, yielding solo-LTR/intact element ratios of 2.5 and 1.2 for Gypsy and Copia retrotransposons, respectively. These results showed an earlier burst of Gypsy retrotransposition than of Copia retrotransposition in the Tu genome; both bursts occurred after the divergence of the A and B genomes[11].

We found substantially higher gene density and recombination rates, as well as lower densities of transposable elements and tandem repeats, near the telomeres of each chromosome (Fig. 1b–l, Extended Data Fig. 3). LTR retrotransposons were distributed unevenly throughout each chromosome. The distribution of Copia elements was enriched at both telomeric–subtelomeric regions, whereas Gypsy retrotransposons were enriched in the pericentromeric–centromeric regions (Fig. 1e, Extended Data Fig. 3). The accumulated gene expression level was higher in subtelomeres than in the centromeric regions (Fig. 1k, Extended Data Fig. 3).

Analyses of genes in the Tu genome together with those from rice[9], maize[12], sorghum[13] and *Brachypodium*[10] were clustered into 24,860 gene families (Extended Data Fig. 4a). Of these, 10,681 families were

Table 1 | Summary of the Tu genome assembly and annotation

Genome assembly	Estimated genome size	4.94 Gb	
	GC content	45.93%	
	N50 length (contig)	344 kb	
	Longest contig	3.00 Mb	
	Total length of contigs	4.79 Gb	
	N50 length (scaffold)	3.67 Mb	
	Longest scaffold	18.76 Mb	
	Total length of scaffolds	4.86 Gb	
	Annotation	**Per cent**	**Total length**
Transposable elements	Retrotransposons	71.83	3.44 Gb
	DNA transposons	7.41	0.35 Gb
	Others	2.19	0.10 Gb
	Total	81.42	3.90 Gb
Protein-coding genes	Predicted genes	41,507	
	Average transcript length	1,453 bp	
	Average coding sequence length	998 bp	
	Average exon length	320 bp	
	Average intron length	508 bp	
	Functionally annotated	36,602	

---

[1]State Key Laboratory of Plant Cell and Chromosome Engineering, Institute of Genetics and Developmental Biology, Chinese Academy of Sciences, Beijing, China. [2]College of Life Sciences, University of Chinese Academy of Sciences, Beijing, China. [3]State Key Laboratory of Plant Genomics, Institute of Genetics and Developmental Biology, Chinese Academy of Sciences, Beijing, China. [4]Keygene N.V., Wageningen, The Netherlands. [5]BGI-Shenzhen, Shenzhen, China. [6]These authors contributed equally: Hong-Qing Ling, Bin Ma, Xiaoli Shi, Hui Liu, Lingli Dong, Hua Sun. *e-mail: hqling@genetics.ac.cn; amzhang@genetics.ac.cn; dwwang@genetics.ac.cn; cliang@genetics.ac.cn

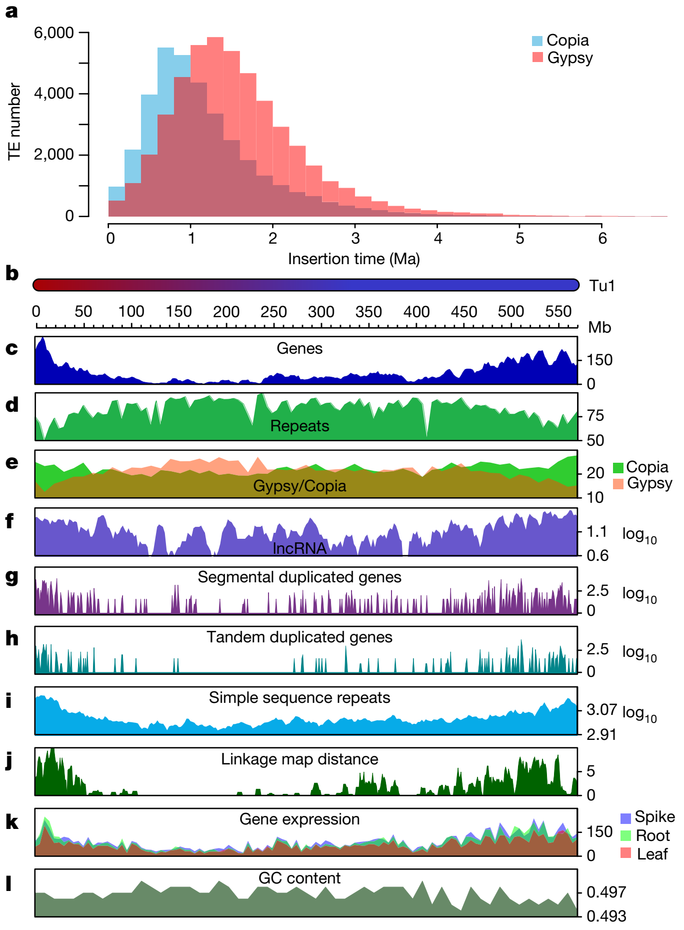


Fig. 1 | Recent LTR retrotransposon bursts in the Tu genome and distribution of genomic components on Tu chromosome 1. a, Insertion burst of LTR retrotransposons of Gypsy and Copia. TE, transposable element. **b**–**l**, Multi-dimensional display of genomic components of Tu chromosome 1. **b**, DNA pseudomolecule. **c**, Gene frequency (number of genes per 10 Mb). **d**, Repeat density (per cent nucleotides per 5 Mb). **e**, Density of LTR retrotransposons (per cent nucleotides per 10 Mb). **f**, Frequency of lncRNA (log[number per 10 Mb]). **g**, Frequency of segmentally duplicated genes (log[number per 1 Mb]). **h**, Frequency of tandemly duplicated genes (log[number per 1 Mb]). **i**, Frequency of simple sequence repeats (log[number per 10 Mb]). **j**, Linkage map distance (cM per 5 Mb). **k**, Accumulated gene expression level ($\log_2$[FPKM (fragments per kilobase of transcript per million mapped reads) per 5 Mb]). **l**, GC content (per cent per 1 Mb).

shared among the five examined plant genomes, representing a core set of genes across these grass genomes. There were 4,610 genes from 1,567 gene families that were specific to Tu, of which many have functional gene ontology annotations relating to responses to stimulus and stress (Extended Data Fig. 4b).

By comparing transcription factors in the Tu genome with those of the six sequenced grass genomes in the iTAK[14] collection, including *Brachypodium*, rice, sorghum, maize, *Aegilops tauschii* (Aet)[15] and Ta[7,16], we found that the number of reproductive meristem (REM) subfamily genes in the transcription factor B3 family[17] was amplified in the genomes of Tu, Aet and Ta (Supplementary Data 3, Extended Data Fig. 4c, Supplementary Information S1.1). The REM subfamily is functionally related to vernalization and flower development[18]. Therefore, we speculate that the amplification of B3 REM transcription factors in wheat genomes might be related to the vernalization process of wheat. Furthermore, we identified 598 disease resistance genes (Supplementary Data 4) and 22 prolamin genes (Supplementary Data 5, Supplementary Information S1.2).

We identified three large structural variations that occurred in either Tu or Ta, with clearly defined boundaries, by comparing the Tu genome to the draft sequences of three sub-genomes of hexaploid wheat[7] (Fig. 2a, b, Extended Data Fig. 5a, Supplementary Information S2.1). We aligned the Tu genome with sequences from six BACs of the A subgenome of *T. turgidum* (Tt) and eleven BACs from the A subgenome of Ta, and found that the unaligned regions between the BAC and the Tu genomic sequences resulted from the insertion of LTR retrotransposons in either Tu or Tt and/or Ta (Extended Data Fig. 5b). Furthermore, we compared the chromosome 7 assembly of the A subgenome of Ta[19] (Ta7A) to Tu chromosome 7 (Tu7), and found that 655 Mb (91.03%) and 536 Mb (90.06%) of Tu7 and Ta7A sequences, respectively, were aligned to each other at a minimum identity of 90% or lower, with many unaligned retrotransposon regions (Extended Data Fig. 5c, d). These results show that the different wheat A genomes experienced large-scale structural rearrangements both before and after polyploidization with other genomes, and have experienced independent gain or loss of LTR retrotransposons after the polyploidization event.

Tu shared with rice, *Brachypodium* and sorghum a common grass ancestor[20] with seven pairs of ancient chromosomes, which became 12 pairs of ancestral chromosomes (still maintained in rice) after one round of whole-genome duplication (WGD, 70 Ma) and two additional chromosomal fusions[21–24]. By studying the collinear relationships among these species (Extended Data Fig. 6, Supplementary Data 6), we found that Tu3 and Tu6 are the most conserved Tu chromosomes, each being derived from a single ancient chromosome shared by Os1, Bd2 and Sb3 and Os2, Bd3 and Sb4, respectively (where Os is *Oryza sativa*, Bd is *Brachypodium distachyon* and Sb is *Sorghum bicolor*). Chromosomes Tu1, Tu2, Tu4 and Tu7 are each composed of chromosomal segments originating from two different ancient chromosomes. Tu5, however, was derived from three ancient chromosomes (Fig. 2c, Supplementary Information S2.2). These results are consistent with the model proposed by Murat et al.[24], and we further narrowed the fusion boundaries to small regions of approximately hundreds of kilobases.

Using an approach based on that described by Murat et al.[25], we inferred 11,718 Tu ancestral grass karyotype (AGK) genes, which account for 31.2% of all chromosome localized Tu genes, slightly lower than the percentage detected in rice (32.4%) and considerably lower than that in *Brachypodium* (47.4%). The AGK genes were depleted in pericentromeric and subtelomeric regions, as well as chromosomal fusion locations (Fig. 2c, Supplementary Information S2.2).

We performed intragenomic comparisons in Tu and found five clearly visible collinear regions in the dot plots (Extended Data Fig. 7, Supplementary Information S2.3). These regions originated from four pairs of anciently duplicated chromosomes[21,22]. However, compared to the rice genome, the collinearity was disrupted between each pair of the Tu chromosome segments derived from the remaining ancient chromosomes, owing to the loss of one or both copies of ancestral genes. For example, of the 2,620 anciently duplicated gene pairs still maintained in rice, approximately 47% and 38% had lost one and both copies, respectively, in the syntenic regions of Tu (Extended Data Fig. 7g).

Comparative analysis of chromosomes Tu3 and Ta3B[26] identified variations between the two chromosomes at both nucleotide and protein levels (Extended Data Fig. 8a–g, Supplementary Information S2.4). We identified 617 Mb (82.6%) and 651 Mb (84.1%) of syntenic sequences in Tu3 and Ta3B, respectively, with only 3,103 (52.32%) genes of Tu3 being aligned to 3,542 (52.99%) genes of Ta3B at a minimum protein identity of 50% and a minimum coverage of 50%. By comparison with syntenic genes from *Brachypodium*, rice and sorghum, we identified 393 and 213 deleted genes and 354 and 648 inserted genes within the collinear segments of Tu3 and Ta3B, respectively (Supplementary Data 7), highlighting the increased number of genes in Ta3B relative to Tu3. We also found a recent LTR retrotransposon burst in Ta3B, which was not observed in Tu3 (Extended Data Fig. 8h). Together, these differences may contribute to the larger size of Ta3B in comparison to Tu3, and suggest that amplification of LTR retrotransposons after the divergence of the A and B genomes may be relevant for wheat genome evolution.

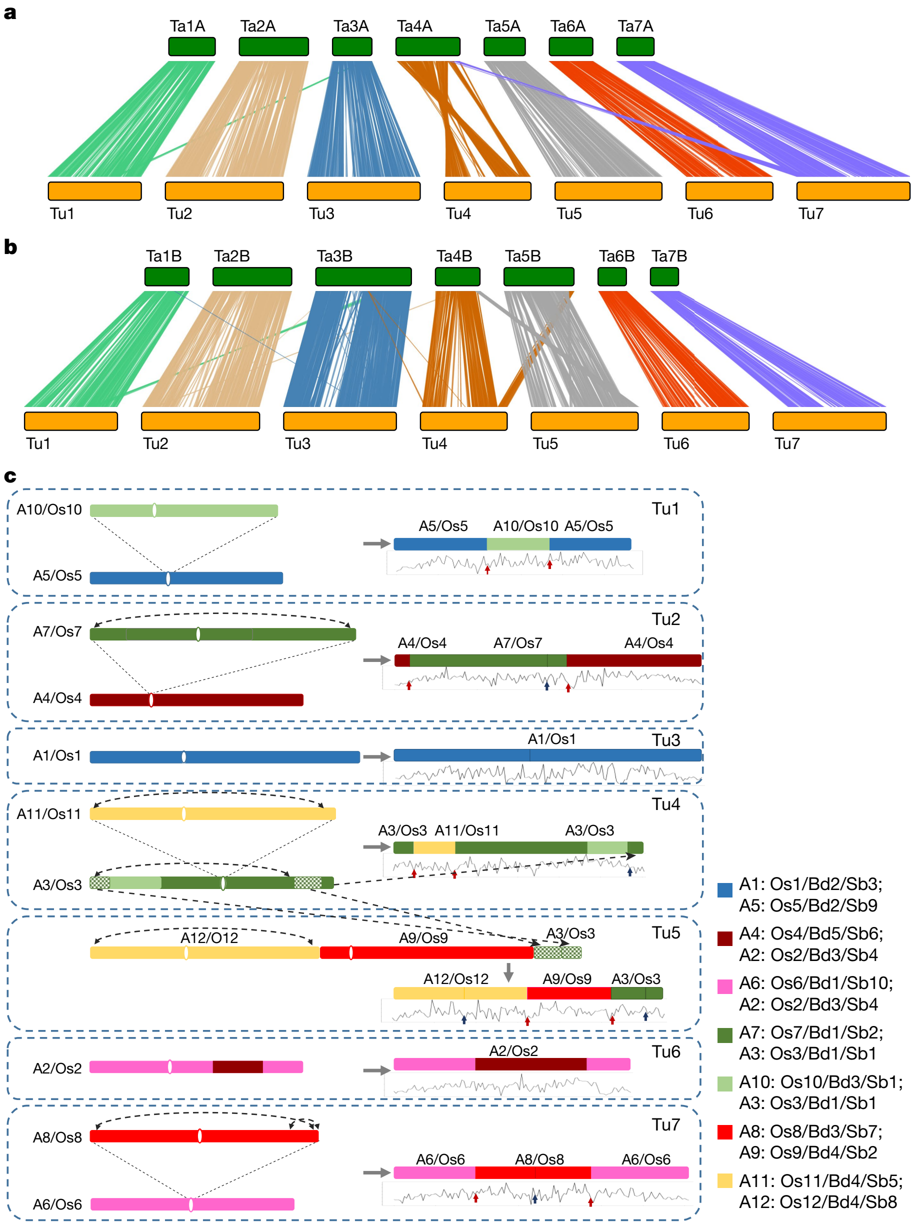


Fig. 2 | Genome synteny to bread wheat (Ta) and an evolutionary model of the Tu chromosomes. **a**, **b**, Synteny of Tu chromosomes with subgenomes A and B of Ta. Each line represents a syntenic block of five or more gene pairs with similarity of 80% or more. Three large structural variations detected are: (1) a reciprocal translocation at the distal end of the long arms between Tu4 and Tu5 that occurred before the polyploidization of the A and B genomes, but after divergence from both B and D genomes; (2) a one-way translocation from Ta7B to Ta4A; and (3) a pericentric inversion on Ta4A involving most of the long and short arms. **c**, Evolutionary model of Tu chromosomes from an ancestral grass genome based on the AGK structure initially defined in Murat et al.[24] and the syntenic relationships of Tu with *B. distachyon* (Bd), rice (Os), and sorghum (Sb). One-directional arrows indicate segment translocations, and bidirectional arrows indicate inversions. Tu1–Tu7, seven chromosomes of Tu; A1–A12, twelve chromosomes of the grass ancestor; Bd1–Bd5, five chromosomes of Bd; Os1–Os12, twelve rice chromosomes; Sb1–Sb10, ten sorghum chromosomes. The seven coloured squares on the right represent seven basic ancient grass chromosomes[24]. The line graphs below Tu chromosomes display the frequency distribution of AGK genes. The red and blue arrows indicate inter- and intra-chromosome fusion locations, respectively, of the ancestral chromosomes in the Tu genome.

For population genetic studies, we sequenced the leaf transcriptome of 147 *T. urartu* accessions collected from six countries in the Fertile Crescent (Armenia, Iran, Iraq, Syria, Turkey and Lebanon) (Fig. 3a, Supplementary Data 8, Supplementary Information S3) and identified 144,806 high-quality SNPs from 22,841 expressed genes (Extended Data Fig. 9a). We analysed population structure using STRUCTURE, and showed that based on this or phylogenetic analyses, the Tu accessions clustered into three groups (Fig. 3b, c). Group I contained 30 accessions from multiple countries. Group II contained 64 accessions, 88% of which were from Lebanon. Group III contained 53 accessions, with 92% from Turkey. These groups have differences in the collection site altitudes, with the majority of Group II accessions being from altitudes above 1,000 m (Extended Data Fig. 9b, Supplementary Data 8). The genetic diversity was lowest in Group II (Extended Data Fig. 9c).

After inoculation with the powdery mildew pathogen *Blumeria graminis* f. sp. *tritici*[27] (Bgt, race E09), 92.2% of the accessions in Group II exhibited resistance, whereas most of the accessions in Groups I and III (96.7% and 90.6%, respectively) were susceptible (Extended Data Fig. 9d). We conducted genomic scans for selective sweeps, and detected 141 ($\pi_{\text{Group I}}/\pi_{\text{Group II}} > 7.7$) and 143 ($\pi_{\text{Group III}}/\pi_{\text{Group II}} > 4.3$) candidate sweep signals based on SNP diversity ratios in the compared groups (Extended Data Fig. 9e). These regions included 239

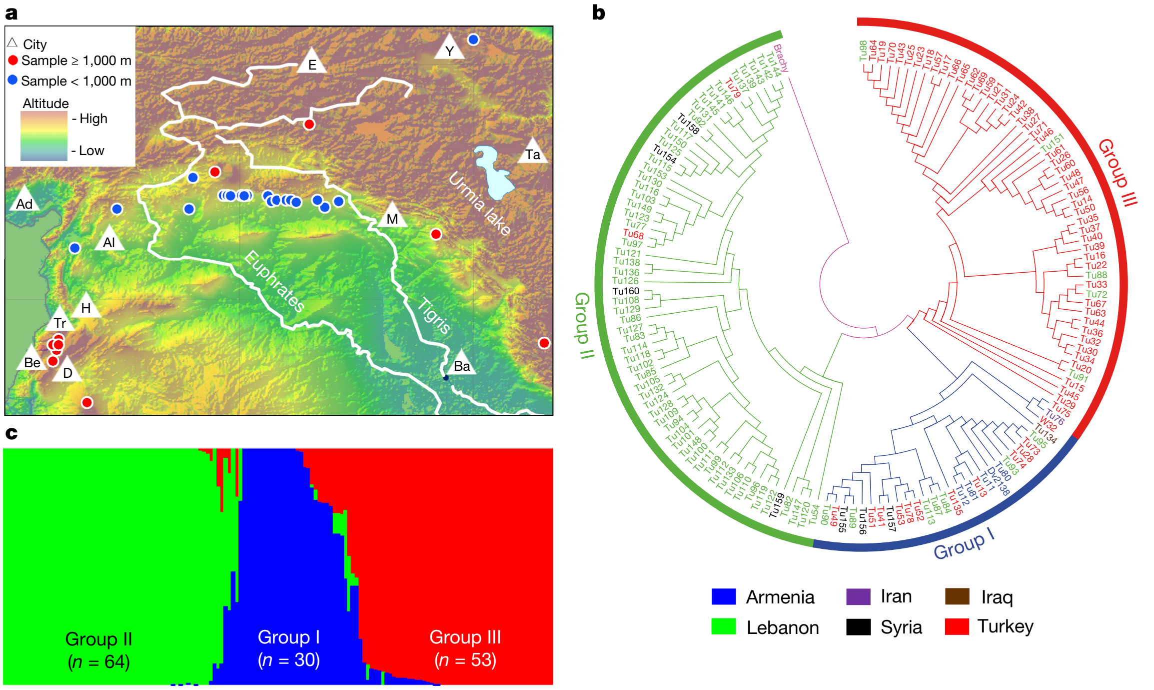


Fig. 3 | Geographic distribution and population structure of Tu. a, Distribution of the 147 Tu accessions at different altitudes of the Fertile Crescent. Also shown are the main land markers, including the Euphrates and Tigris rivers, Urmia lake and 11 cities (Adana (Ad), Aleppo (Al), Baghdad (Ba), Beirut (Be), Damascus (D), Erzurum (E), Homs (H), Mosul (M), Tabriz (Ta), Tripoli (Tr) and Yerevan (Y)). The map was drawn using the online mapping tool ArcGIS (version 10.1, www.esri.com). **b**, Phylogenetic clustering of the 147 accessions into three groups, with *B. distachyon* as the outgroup. **c**, Population structure analysis of Tu accessions, clustering with three groups that are similar to those from the phylogenetic analysis shown in **b**.

high-confidence predicted genes (Supplementary Data 9), including those with functional annotations for transcriptional regulation (32 genes), signal transduction (14) or detoxification of reactive oxygen species and stress defence (14); 23 of the genes were Tu-specific. Previous studies have found that plant and animal adaptation to high altitudes involves genes with functions in diverse physiological and molecular processes[28,29]. Further analysis showed a wall-associated receptor protein kinase gene (*TuWAK*, TuG1812G0400002796) within a selective sweep signal, which had two haplotypes (Hap1 and Hap2) with differing distributions among the three groups (Extended Data Fig. 9f, g). Hap1 was the major haplotype in Group II, and associated with resistance to Bgt. Hap2 was the main haplotype in Groups I and III, and associated with Bgt susceptibility. The maize homologue *ZmWAK* is involved in defence against fungal pathogens[30]. Therefore, *TuWAK* may have been under selection in Group II accessions, and contributed to both high altitude adaptation and powdery mildew resistance.

Online content

Any Methods, including any statements of data availability and Nature Research reporting summaries, along with any additional references and Source Data files, are available in the online version of the paper at https://doi.org/10.1038/s41586-018-0108-0.

Received: 11 September 2016; Accepted: 29 March 2018;
Published online 9 May 2018.

1. Dvorák, J., Terlizzi, P., Zhang, H. B. & Resta, P. The evolution of polyploid wheats: identification of the A genome donor species. *Genome* **36**, 21–31 (1993).
2. Peng, J. H., Sun, D. H. & Nevo, E. Domestication evolution, genetics and genomics in wheat. *Mol. Breed.* **28**, 281–301 (2011).
3. Ferrarini, M. et al. An evaluation of the PacBio RS platform for sequencing and de novo assembly of a chloroplast genome. *BMC Genomics* **14**, 670 (2013).
4. Zheng, G. X. et al. Haplotyping germline and cancer genomes with high-throughput linked-read sequencing. *Nat. Biotechnol.* **34**, 303–311 (2016).
5. Lam, E. T. et al. Genome mapping on nanochannel arrays for structural variation analysis and sequence assembly. *Nat. Biotechnol.* **30**, 771–776 (2012).
6. Ling, H.-Q. et al. Draft genome of the wheat A-genome progenitor *Triticum urartu*. *Nature* **496**, 87–90 (2013).
7. International Wheat Genome Sequencing Consortium (IWGSC). A chromosome-based draft sequence of the hexaploid bread wheat (*Triticum aestivum*) genome. *Science* **345**, 1251788 (2014).
8. Liang, C., Mao, L., Ware, D. & Stein, L. Evidence-based gene predictions in plant genomes. *Genome Res.* **19**, 1912–1923 (2009).
9. International Rice Genome Sequencing Project. The map-based sequence of the rice genome. *Nature* **436**, 793–800 (2005).
10. International Brachypodium Initiative. Genome sequencing and analysis of the model grass *Brachypodium distachyon*. *Nature* **463**, 763–768 (2010).
11. Marcussen, T. et al. Ancient hybridizations among the ancestral genomes of bread wheat. *Science* **345**, 1250092 (2014).
12. Schnable, P. S. et al. The B73 maize genome: complexity, diversity, and dynamics. *Science* **326**, 1112–1115 (2009).
13. Paterson, A. H. et al. The *Sorghum bicolor* genome and the diversification of grasses. *Nature* **457**, 551–556 (2009).
14. Zheng, Y. et al. iTAK: A program for genome-wide prediction and classification of plant transcription factors, transcriptional regulators, and protein kinases. *Mol. Plant* **9**, 1667–1670 (2016).
15. Jia, J. et al. *Aegilops tauschii* draft genome sequence reveals a gene repertoire for wheat adaptation. *Nature* **496**, 91–95 (2013).
16. Brenchley, R. et al. Analysis of the bread wheat genome using whole-genome shotgun sequencing. *Nature* **491**, 705–710 (2012).
17. Swaminathan, K., Peterson, K. & Jack, T. The plant B3 superfamily. *Trends Plant Sci.* **13**, 647–655 (2008).
18. Levy, Y. Y., Mesnage, S., Mylne, J. S., Gendall, A. R. & Dean, C. Multiple roles of *Arabidopsis* VRN1 in vernalization and flowering time control. *Science* **297**, 243–246 (2002).
19. Clavijo, B. J. et al. An improved assembly and annotation of the allohexaploid wheat genome identifies complete families of agronomic genes and provides genomic evidence for chromosomal translocations. *Genome Res.* **27**, 885–896 (2017).
20. Kellogg, E. A. Evolutionary history of the grasses. *Plant Physiol.* **125**, 1198–1205 (2001).
21. Wang, X., Shi, X., Hao, B., Ge, S. & Luo, J. Duplication and DNA segmental loss in the rice genome: implications for diploidization. *New Phytol.* **165**, 937–946 (2005).
22. Salse, J. et al. Identification and characterization of shared duplications between rice and wheat provide new insight into grass genome evolution. *Plant Cell* **20**, 11–24 (2008).
23. Singh, N. K. et al. Single-copy genes define a conserved order between rice and wheat for understanding differences caused by duplication, deletion, and transposition of genes. *Funct. Integr. Genomics* **7**, 17–35 (2007).
24. Murat, F. et al. Shared subgenome dominance following polyploidization explains grass genome evolutionary plasticity from a seven protochromosome ancestor with 16K protogenes. *Genome Biol. Evol.* **6**, 12–33 (2014).

25. Murat, F., Armero, A., Pont, C., Klopp, C. & Salse, J. Reconstructing the genome of the most recent common ancestor of flowering plants. *Nat. Genet.* **49**, 490–496 (2017).
26. Choulet, F. et al. Structural and functional partitioning of bread wheat chromosome 3B. *Science* **345**, 1249721 (2014).
27. Zhang, J. et al. Coexpression network analysis of the genes regulated by two types of resistance responses to powdery mildew in wheat. *Sci. Rep.* **6**, 23805 (2016).
28. Ai, H. et al. Population history and genomic signatures for high-altitude adaptation in Tibetan pigs. *BMC Genomics* **15**, 834 (2014).
29. Ma, L. et al. Physiological, biochemical and proteomics analysis reveals the adaptation strategies of the alpine plant *Potentilla saundersiana* at altitude gradient of the Northwestern Tibetan Plateau. *J. Proteomics* **112**, 63–82 (2015).
30. Zuo, W. et al. A maize wall-associated kinase confers quantitative resistance to head smut. *Nat. Genet.* **47**, 151–157 (2015).

Acknowledgements The authors thank M.-C. Luo and J. Dvorak (UC Davis) for providing the leaf material of G1812 for constructing BAC libraries. This work was supported by grants from the Chinese Academy of Sciences (QYZDJ-SSW-SMC001, XDA08010404) and by grants from the Ministry of Science and Technology of China (2016YFD0101004, 2010DFB33540, 2012AA10A308).

Reviewer information *Nature* thanks M. D. Clark and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions H.-Q.L., C. L., A.Z., Z.L. and D.W. were responsible for project initiation. Project coordination was by H.-Q.L. and C.L. The project was managed by H.-Q.L, C.L., B.Z. and W.Z. Data generation and analysis were performed by B.M., X.S., H.L., L.D., H.S., Y.Ca., S.Zhe., Ya.L., Ye.L., Y.Y., Q.G., H.D., M.Q., Y.Cu., H.Y., N.W., C.C., H.W., Y.Z., J.Z., Yiw.L., W.H., S.Zha., M.J.T.V.E., J.T. and H.M.A.W. Experiments and analyses were designed by H.-Q.L, C.L., D.W., X.S., B.M., L.D. and H.S. The paper was written by H.-Q.L., C.L., X.S., B.M., L.D., A.Z. and D.W. All authors read and commented on the manuscript.

Competing interests The authors declare no competing interests.

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0108-0.

Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0108-0.

Reprints and permissions information is available at http://www.nature.com/reprints.

Correspondence and requests for materials should be addressed to H.-Q.L. or A.Z. or D.W. or C.L.

METHODS

Plants. The accession G1812 (PI428198) of *T. urartu* was previously shown by restriction fragment length polymorphism (RFLP) analysis to have the closest relationship to the A subgenome of hexaploid wheat[31]. Its genome was sequenced using a whole-genome shotgun strategy on the Illumina HiSeq2000 platform, and a draft genome has been generated[6]. In this study, we also used accession G1812 of *T. urartu* to improve its genome assembly quality. We combined BAC-by-BAC sequencing with single-molecule real-time (SMRT) sequencing technology (Pacific Biosciences), and new mapping technologies (BioNano genome map and 10× Genomics linked reads) to generate a high-quality reference sequence of *T. urartu*.

BAC library construction. The genomic BAC libraries of *T. urartu* accession G1812 (PI428198) were constructed by AMPLICON Express (Pullman). In brief, nuclei were isolated from leaves of one-month-old plants, and embedded in agarose plugs. After lysis of nucleic membrane and digestion of proteins, intact high molecular weight (HMW) genomic DNA was extracted as described[32]. The HMW DNA was partially digested using EcoRI, HindIII and MboI to generate insertion DNA fragments for BAC library construction. Subsequently, the DNA fragments (average sizes 120–180 kb) were ligated into the appropriate sites on the vector pCC1BAC or pECBAC1 or pIndigoBAC-5. The ligations were transformed into *Escherichia coli* DH10B cells (phage resistant). Transformants were robotically picked and arrayed onto 384-well plates. All plates were assigned a barcode and recorded in a database. In total, 470,016 BAC clones were obtained, including 184,320 clones in the EcoRI-digest library with an average insert size of 125 kb, 193,536 clones in the HindIII-digest library with an average insert size of 110 kb, and 92,160 clones in the MboI-digest library with an average insert size of 115 kb. The total insert DNA length was about 54.9 Gb, approximately 11 equivalents of the whole genome of *T. urartu*.

Whole-genome profiling of *T. urartu*. To generate BAC fingerprint contigs, 451,584 BAC clones (179,712 EcoRI clones, 179,712 HindIII clones and 92,160 MboI clones) representing about 10 equivalents of the *T. urartu* genome were analysed using the whole-genome profiling (WGP) technology by Keygene N.V.[33]. After pooling of individual BAC clones, isolation of pooled BAC DNA, digestion with HindIII/MseI, ligation of Illumina Hiseq adaptor sequences and sequencing from the HindIII side using Illumina HiSeq with 100-nt read length, we obtained 1,108,197 filtered WGP tags and 345,233 tagged BACs. Then, the sequence-based physical maps were assembled using an adapted version[33] of the FingerPrint Contig (FPC) software[34] to generate a high-stringency map (using a threshold of 1.0×10^{-28} DQ) and a reduced-stringency map, in which end-to-end merging of contigs was done and singletons were added in a number of repetitive steps (1.0×10^{-25} DQ + end merges 10^{-15} + singleton merges 10^{-15}). The total length of BAC contigs reached 5.52 Gb with a BAC contig N50 size of 340 kb for high-stringency WGP map assembly and 4.68 Gb with a BAC contig N50 size of 656 kb for reduced-stringency WGP map assembly.

Selection of BAC clones and BAC DNA extraction. Using WGP data, 47,200 BAC clones were selected from the tagged 345,233 BAC clones, which were assembled on 20,702 BAC contigs with high-stringency WGP assembly, according to a minimal tiling path (MTP) principle for genomic sequencing via BAC-by-BAC sequencing. The selected MTP BAC clones were divided into two groups (the two neighbouring BACs on each FPC were separated into group A and group B to avoid the misdistribution of sequence reads on BACs after sequencing).

For BAC DNA extraction, each selected BAC clone was incubated in 400 μl lysogeny broth (LB) with 12.5 μg/ml chloramphenicol on 96-well plates at 400 r.p.m. and 37 °C for 16 h. After that, we pooled 2,304 individual BAC clones from 24 pieces of 96-well plates into 96 pools using 2D pooling strategy at horizontal and vertical levels (only BACs from the same group could be pooled together). Each pool contained 48 different BAC clones. For the pooling, 75 μl bacterial solution of each BAC clone was picked up and pooled together as described above for BAC DNA extraction. The BAC DNA was extracted using PhasePrep BAC DNA Kit (Sigma-Aldrich) following the manufacturer's instructions with some modifications. In brief, bacterial cells were collected by centrifugation for 10 min at 4 °C, and re-suspended in 120 μl chilled resuspension solution containing RNase A solution (20 mg/ml). After addition of 120 μl lysis solution, the bacterial solution was well mixed with the lysis solution by gentle turnover four times, and put on ice for 4 min. Subsequently, 120 μl pre-chilled neutralization solution was added and mixed again by gentle turnover four times. After incubation on ice for 5 min, the lysed bacterial solution was centrifuged at 13,000 r.p.m. and 4 °C for 4 min. Then, the supernatant was transferred to a clean 1.5-ml Eppendorf tube and mixed with 250 μl isopropanol by gentle turnover several times. BAC DNA was pelleted by centrifugation at 13,000 r.p.m. and 4 °C for 20 min. After removal of the supernatant, the pellet was washed using 100 μl 70% ethanol and dried in vacuum. To remove RNAs, we added 500 μl elution solution and 1 μl tenfold diluted RNase cocktail to suspend the pellet. After incubation at 60 °C for 5 min, 20 μl sodium acetate buffer solution and 50 μl endotoxin removal solution were added separately to the DNA solution, mixed and incubated at 37 °C for 5 min. Next, the clean upper phase solution was transferred to a new Eppendorf tube and mixed with 270 μl DNA precipitation solution, and BAC DNA was pelleted by centrifugation at 13,000 r.p.m. and 4 °C for 10 min. Then, the DNA pellet was washed twice using 400 μl 70% ethanol and dried in vacuum. Finally, the BAC DNA was dissolved in 25 μl sterile deionized water, and kept at −20 °C until use. In total, 47,223 BAC clones were pooled into 984 BAC pools and their DNAs were extracted and used for sequencing

Preparation of Illumina sequencing libraries and sequencing. For sequencing of BACs by Illumina HiSeq2500, we constructed paired-end libraries with ~300 bp insert size following the protocol of NEBNext Ultra DNA Library Prep Kit for Illumina (New England Biolabs) with a slight modification. In brief, 50 ng DNA of each BAC pool (containing 48 BAC clones) was sheared in a Covaris S2 focused ultrasonicator to an average insert size of 300 nt. After end reparation with the NEBNext Ultra End Repair/dA-Tailing Module Kit (New England Biolabs), the DNA fragments were ligated with barcode adaptors, and 96 BAC DNA samples with different barcode adapters were mixed together and purified using a QIAquick PCR Purification Kit (Qiagen). Subsequently, approximately 300-bp DNA fragments were selected again on 2% agarose gel, and amplified by PCR with 11 cycles. The library was then sequenced using 150 base-length read chemistry in a paired-end flow cell on the Illumina HiSeq2500 after library profile analysis by the Agilent 2100 Bioanalyzer and qPCR quantification. Paired-end sequencing was performed following the manufacturer's protocol (http://www.illumina.com/) based on the workflow: cluster generation, template hybridization, isothermal amplification, linearization, blocking, denaturation and hybridization of sequencing primers. The base-calling pipeline (Hiseq2500) was used to detect bases from the raw fluorescent images. In total, 39 libraries, including 2,347 pools (48 BACs per pool) with 300 bp insert size, were prepared, 33 sequencing lanes were run in Hiseq2500 and 2,102 Gb of raw sequence data was generated (Extended Data Fig. 1b).

Data quality control. The raw datasets were first filtered by trimming reads with low-quality bases (quality < 2) at the front end and reads with quality <10 at the back end, and by discarding reads with 20% low-quality bases (quality <10) and reads with a length <75 base pairs. Then, we removed the contaminated sequence reads by blasting the sequence reads with the genomes of *E. coli*, mitochondria, chloroplast, and the vector sequence as well human genomic sequence with BWA software[35] (Burrows-Wheeler Aligner, http://bio-bwa.sourceforge.net/). On average, 20% of reads were aligned to microbial genome; 4.5% of reads were aligned to vector genome; and 0.5% of reads were aligned to chloroplast, mitochondrion and human genomes. Finally, 75% of sequence reads with a total length of 1,471 Gb (about 294× *T. urartu* genome) remained for assembly of the *T. urartu* genome.

To confirm that the sequence reads assigned to each BAC were from the right BAC clone, we blasted sequence reads of each BAC pool against the WGP BAC tags. BAC pools with sequence reads less than 400 Mb were sequenced again to obtain enough sequences for assembly. We also found that sequence reads of a few BAC pools were blasted to two or more neighbour WGP BAC tags owing to BAC clone contamination. Thus, these BAC clones were picked again and their DNA was extracted and resequenced.

Preparation of PacBio sequencing libraries and sequencing. To enable assembly of complex repeat structure and GC- and AT-rich regions, which are often unassembled or highly fragmented in next generation sequencing (NGS)-based draft genomes, we also performed whole-genome shotgun sequencing using SMRT sequencing technology (Pacific Biosciences). The library preparation and sequencing were done by Nextomics. Sequencing libraries with 20-kb DNA inserts were prepared following the protocol of the PacBio template preparation kit (DNA Template Prep Kit 1.0) and sequenced using Pacific Biosciences RSII instrument. A total of 109 SMRT cells were processed. Subread filtering was performed using Pacific Biosciences SMRT analysis software (v2.3.1) with the parameters (subread length = 50, minimum polymerase read quality = 75, minimum polymerase read length = 50). In total, 97 Gb clean sequence data were obtained with an average subread length of 8.1 kb and an N50 subread length of 11.2 kb (Extended Data Fig. 1b).

Owing to the high error rate of SMRT reads, we constructed a PCR-free paired-end library with 500 bp insert size using the whole-genome DNA of *T. urartu* with the PCR-free protocol (Illumina kit FC-121-3001) and sequenced it by Hiseq2500 with two lanes. In total, 130 Gb whole genome shotgun paired-end reads with 250 bp read length were obtained (Extended Data Fig. 1b). Subsequently, we filtered the low quality reads and contamination reads with bacterial genome and vectors and obtained 107 Gb (21×) clean reads, which was used for error correction of the SMRT reads.

BAC assembly. For assembly, the pipeline flowchart outlined in Extended Data Fig. 1a was followed. First, the Illumina clean reads in each BAC pool were separately assembled into contigs using MaSuRCA software[36]. Then, Illumina clean reads in each vertical BAC pool were aligned against the contigs in all horizontal BAC pools with BLAT[37] and vice versa. The reads, which had ≥90% coverage and ≥99% identity, and appeared only once in all crossing BAC pools, were selected as input reads to the BAC at the cross point. Those input reads assigned to each BAC were assembled again using MaSuRCA software to obtain sequence contigs of each individual BAC. At this stage, the total contig length of each BAC reached 125 kb and the contig N50 was 35 kb on average.

Next, 107 Gb of PCR-free Illumina clean reads with 250 bp read length was used to correct the 97 Gb (19.5×) PacBio raw reads using Proovread[38], yielding 72 Gb (14×) of corrected PacBio reads. To fill gaps between contigs in each BAC, the corrected PacBio reads were aligned to the BAC contigs with BLASR[39] (parameters identity = 95%, minlength = 1 kb). Subsequently, the sequence contigs of each BAC were connected with the best-aligned PacBio reads using a customized perl script. After this process, 46,374 of 47,223 BACs (98.2%) were assembled into a single contig and 610 BACs into two contigs, and more than 99.4% of BACs showed an assembled sequence length larger than 100 kb. Finally, we connected the BAC sequences iteratively into contigs based on their overlapping relationship within FPC contig, and obtained 5.33 Gb in total length with a contig N50 of 183 kb.

Assembly of missing regions. Owing to the cleavage bias of restriction enzymes and low genome coverage (8×) of the BAC libraries used for FPC construction, some regions of the genome were missed from the BAC sequences. To assemble the missed regions, we tried to assemble the whole genome using the corrected PacBio reads with Celera Assembler. However, the Celera Assembler encountered an error during the assembly. The main reason might be the low sequencing depth of PacBio reads for such a complex genome. Therefore, we retrieved the corrected SMRT reads and the previously reported sequence contigs of *T. urartu*[6], which were not covered by the BAC sequences at a minimum sequence identity of 95% and 98%, respectively, for assembly. We assembled them using MaSuRCA[40] with default parameters. In total, we assembled additional 204 Mb of sequences into contigs, and added them to the genome assembly.

Scaffolding using BioNano genome map and 10 × genomics linked reads. For construction of a BioNano genome map, 10-day-old seedlings of G1812 were harvested. The DNA isolation, sequence-specific labelling of megabase gDNA for Irys mapping by nicking, labelling, repairing, and staining (NLRS) and chip analysis were performed by Genergy Bio Technology according to the manufacturer's instructions (BioNano Genomics). In brief, the enzyme Nt.BspQI with an appropriate label density (11.5 labels per 100 kb) was selected and applied to digest long-range DNA fragments. Then, the NLRS number per DNA fragment was determined using the BioNano Irys system. In total, 502 Gb BioNano mapping molecules with an average length of 265.71 kb was collected. After filtering the molecules with a cutoff at a minimum length of 150 kb and 8 labels per molecule, 417 Gb BioNano molecules (83× effective depth) with an average length of 294.95 kb was obtained (Extended Data Fig. 1b). Furthermore, we used autonoise[5] and other default parameters in IrysSolve tools based on *T. urartu* genome sequences to determine the de novo assembly noise. The RefAligner and Assembler programs in IrysSolve tools were used to assemble these BioNano molecules with initial assembly P value of 1×10^{-10} and extension/refinement P value of 1×10^{-11}. After the aforementioned processes, 9,112 BioNano genome maps with a total map length of 4.68 Gb were generated. The N50 length of the BioNano genome map was 0.61 Mb and the average map length was 0.54 Mb.

We used BioNano genome maps and the *T. urartu* sequence contigs to generate hybrid maps with initial and final alignment P value of 1×10^{-10}, chimaeric/conflicting P value of 1×10^{-13} and merging P value of 1×10^{-11}. The sequence contigs that had conflicts with BioNano genome maps were cut into sub-sequences for additional hybrid map generation. We identified only 631 chimaeric contig assemblies. Finally, hybrid sequence scaffolds were generated based on these hybrid genome maps. The overlapping contigs in hybrid scaffolds were aligned using MUMmer[41] and merged into sequence contigs.

For generation of 10× Genomics linked reads, the leaf DNA of G1812 was extracted from 10-day-old seedlings and used for the construction of 10× Genomics libraries following the manufacturer's protocol (10× Genomics). Then, we used the chromium system to barcode short fragments onto long DNA molecules ($\geq$50 kb), and HiSeq X Ten to sequence these short fragments into standard Illumina paired-end reads as linked-reads. Short reads from the same long DNA molecule shared the same barcode. In total, 57 Gb Illumina paired-end reads (11× effective depth) were produced. Furthermore, we used the software longranger-2.1.2[4] to align the linked reads to the scaffolds of *T. urartu*. A mean molecule length of 31.39 kb and mean of 21 linked reads per molecule were obtained. Subsequently, the large_sv_call subprogram in longranger-2.1.2[4] was used to find connection information for scaffolds in the *T. urartu* genome. After this step, a large number of sequence scaffolds and contigs that were not on BAC FPC and were not connected by BioNano genome maps owing to their short length were connected to longer scaffolds (>100 kb).

Additionally, the previously reported 10-kb and 20-kb mate pair reads[6] were also used to connect sequence contigs into scaffolds using SSPACE[42] with default parameters. To minimize errors introduced in the scaffolding step, we only connected the sequence contigs supported by linkage evidence from BioNano and 10× Genomics. The SMRT raw reads were also applied to fill the gap in scaffolds using PBJelly[43] with default parameters to generate longer sequence contigs.

Generation of pseudomolecules. To anchor the assembled sequences onto chromosomes, we developed an F2 population containing 475 individuals from a cross between accessions G1812 and G3146. We sequenced this population using the restriction enzyme TaqI-associated DNA sequencing (RAD-seq) method for calling SNPs and constructing a genetic map. In total, 981 Gb Illumina sequences was generated (1,028 Mb per F2 individual on average). These Illumina reads were mapped to our assembled sequences using BWA MEM[35], and SNPs were called using the GATK pipeline[44]. In total, 3,751,342 SNPs were identified between G1812 and G3146. The SNPs were then filtered using parameters DP (depth) $\geq$2 and MAF (minor allele frequency) $\geq$ 0.1. Finally, 430,979 high quality SNPs were selected and used for genotyping the sequenced F2 individuals. Adjacent SNPs were merged together using a sliding window method (window:50 SNPs; step:50 SNPs) into bins. The bins as input markers were grouped into seven linkage groups through the ML (maximum likelihood) algorithm in joinMAP 4.1[45], and the linkage map of bins was created using the Kosambi model in MSTmap[46]. In total, 22,386 bins were anchored on the seven chromosomes of *T. urartu*. Using the SNP bin markers, 27,587 sequence contigs were anchored onto chromosomes (Supplementary Data 1). We then assigned the assembled scaffolds onto their corresponding positions on the chromosome using the mapping information from SNPs, and generated seven pseudomolecules. Subsequently, adjacent bins with the same genotype within a physical distance of 100 kb were merged into larger bins to remove bins with spurious or missing genotypes. Furthermore, we set an upper limit on the physical distance between two adjacent crossovers of at least 1 Mb, and allowed only at most ten recombination events on each chromosome, and changed the incompatible genotypes that caused spurious double crossover events to missing data. Finally, we recalculated the genetic distances from the recomputed recombination frequencies through the Kosambi mapping function for each linkage group, and obtained a genetic map consisting of 4,506 high-confidence bins. The accumulated genetic linkage distance of the genetic map was 1,444 cM (Extended Data Fig. 1e).

Assembly evaluation. To evaluate the quality of our assembly, we compared the Tu genome to 12 previously published BAC sequences (Extended Data Fig. 2a) from the *T. urartu* G1812 genome downloaded from NCBI (http://www.ncbi.nlm.nih.gov/nuccore/?term=Triticum+urartu+BAC) using NUCmer (parameter: -mum -mincluster 700) in MUMmer package[41]. Then we drew the dot plot using mummerplot in the same package with default parameters. We identified the repeat sequences using BLAST with E value 1×10^{-10} against the TREP database (http://botserv2.uzh.ch/kelldata/trep-db/index.html) and PGSB Repeat Element Database (http://pgsb.helmholtz-muenchen.de/plant/recat/) (Extended Data Fig. 2). All of the BAC sequences were nearly completely covered by our assembled pseudomolecules with an average coverage of 98.89% and sequence identity of 99.66%. These results indicate that we have generated a high quality genome sequence of *T. urartu*.

To evaluate the per base error rate, the WGS short reads, including 'WGS HiSeq PCRFree 2 × 250' (Extended Data Fig. 1b) and 'WGS Hiseq 2 × 150 (GenBank accessions SRR124016 ~ SRR124023)', were aligned to the *T. urartu* genome with BWA MEM. After filtering out the short alignments (<50 bp), we found that 98.56% of the *T. urartu* genome was covered. Meanwhile, 98.45% of reads in 'WGS HiSeq PCRFree 2 × 250' and 98.78% of 'BGI Hiseq 2 × 150' reads were aligned to the *T. urartu* genome. After removing the aligned reads with identity $\leq$98%, we used GATK to call variants from the data above and obtained 541,849 SNPs (0.011%, one per 9 kb) and 128,592 indels with a total size of 281,155 bp (0.006% base error).

To evaluate structural chimeric error rate, we compared the *T. urartu* genome contigs and the BioNano genome map by RefAligner program in IrisSolve package from BioNano Genomics (https://bionanogenomics.com/support/software-downloads/), and detected 5,346 collapse/expansion regions (>1 kb) using the SV detect program in the same package. These collapse/expansion regions covered 54.92 Mb (1.13%) of the assembled genome (4.86 Gb), with a maximum length of 124.43 kb, a minimum length of 1,012 bp, and an average length of 10.27 kb. Among them, there were 4,971 collapsed regions covering 51.26 Mb (1.05% of the assembled genome 4.86 Gb) with a maximum length of 124.43 kb, a minimum length of 1,018 bp and an average length of 10.31 kb, and 375 expansions covering 3.67 Mb (0.07% of total assembled genome 4.86 Gb) with a maximum length of 62.36 kb, a minimum length of 1,012 bp and an average length of 9.78 kb. All the data strongly indicate that the quality of our genome assembly is high and reliable.

Annotation and analysis of repetitive elements. Repetitive sequences and transposable elements in the *T. urartu* genome were identified using a combination of ab inito and homology-based methods at both DNA and protein levels. In brief, an ab inito repeat library for *T. urartu* was predicted using LTR_FINDER v1.0.2[47], RepeatModeler (v1.0.3) with default parameters. The library was aligned to PGSB Repeat Element Database (http://pgsb.helmholtz-muenchen.de/plant/recat/) to classify the type of each repeat family. For identification of the repeats throughout the genome, RepeatMasker (v3.2.9) was applied with both the ab initio repeat databases and Repbase (http://www.girinst.org/repbase) using the WU-BLASTX search engine. Overlapping transposable elements belonging to the same repeat class were collated and combined. In addition, we annotated the tandem repeats using the software Tandem Repeats Finder (TRF, v4.04)[48].

Solo-LTR was identified using LTRharvest[49]. A customized script was implemented to identify intact LTR/Copia and LTR/Gypsy retrotransposons. The ClustalW program[50] was applied to align 5′ and 3′ solo-LTRs to intact LTR elements. The evolutionary distance of the two LTR sequences was estimated using the Kimura two-parameter method embedded in baseml program in PAML[51]. A substitution rate of 1.3×10^{-8} mutations per site per year was used to convert evolutionary distance between 5′ and 3′ solo-LTRs to insertion age of retrotransposons[52]. In total, we identified 35,559 and 48,370 intact Copia and Gypsy retrotransposons, respectively. To identify solo-LTRs, we first excluded intact LTR transposable elements from the dataset of Gypsy and Copia LTR-retrotransposons, and aligned all known LTR segments to those damaged LTR-retrotransposons. The transposons that were similar (identity >85%) to a known LTR segment were identified as solo-LTRs. We verified that only 2.28% of our identified LTRs overlapped with BioNano collapse/expansion regions and 1.76% overlapped with PacBio junction regions, indicating that very few LTRs were affected by misassembly of sequences.

Simple sequence repeat (SSR) markers are useful in plant genetic analysis. Therefore, we detected SSR markers within our assembled sequences using MISA software (http://pgrc.ipk-gatersleben.de/misa/misa.html). A total of 486,506 SSRs were identified. Dinucleotide was the most common repeat motif with a frequency of 35.6% (173,125 SSRs), followed by mono- (170,845, 35.1%), tri- (88,008, 18.1%), and hexa-nucleotide (7,145, 1.5%) repeat motifs. The distribution of SSRs is shown on Fig. 1i and Extended Data Fig. 3.

Annotation and analysis of non-coding RNAs. Non-coding RNAs of *T. urartu*, including rRNAs, tRNAs, miRNAs and snoRNAs, were analysed. We used tRNAscan-s.e. (version 1.23) with eukaryote parameters to predict tRNAs[53]. The miRNA and snoRNA genes were predicted using Infernal software (version 1.0)[54] to search the genome against the Rfam database (http://rfam.xfam.org/, release 9.1) with default parameters. The rRNA sequences were predicted using BLASTN (E value $<1 \times 10^{-5}$) to align the known rRNA genes (5S, 5.8S, 18S, and 28S) of both *T. aestivum* and *Arabidopsis* from GenBank to the draft genome. Additionally, lncRNAs were identified with rigorous criteria: (1) transcript length must be longer than 200 bp; (2) transcript must contain no open reading frame (ORF) longer than 50 amino acids; (3) the Coding Potential Calculator (CPC)[55] was used to predict the coding potential of each transcript, and those with CPC scores >0 were discarded.

Gene prediction and functional annotation. We predicted a set of genes in the *T. urartu* genome using evidence-based Gramene pipeline[8] by combining protein, cDNA, EST and RNA-seq evidence. We downloaded 559,967 mRNAs (https://www.ncbi.nlm.nih.gov/nuccore/?term=Triticum+urartu) and 1,283,261 ESTs (https://www.ncbi.nlm.nih.gov/nucest/?term=Triticum) of *T. urartu* from NCBI nucleotide database as same-species cDNAs and same-species ESTs. SwissProt proteins for plants were cleaned up by removing redundant sequences with a minimum threshold of 80% for both identity and coverage, which left us 340,312 sequences as protein evidence. The protein evidence also included 1,795 wheat proteins downloaded from GenBank. The mRNAs and ESTs of monocot species other than wheat were used as cross-species evidence; these were downloaded from NCBI and filtered to remove redundant sequences with a cutoff of 90% for both identity and coverage, which resulted in 548,604 cDNAs and 978,696 ESTs. RNA-seq data from 243 samples (2.47 Tb) of *T. urartu* and bread wheat were downloaded from NCBI and were assembled into contigs using SOAPdenovo-trans v1.03 (http://soap.genomics.org.cn/SOAPdenovo-Trans.html). The assembled contigs were used as same-species EST evidence.

The transcripts predicted from each evidence type were combined using the Gramene pipeline to generate 115,255 potential genes as a raw set, of which the majority were spurious genes with only EST or RNA-seq evidence support. The expression value of each gene in 243 wheat RNA-seq samples was calculated using Cufflinks[56] (-u). The raw gene set was filtered to generate a core set of 41,507 protein-coding genes by removing transposable element-related genes, pseudogenes and non-coding genes. We simply define a gene to be transposable element-related if its protein has >50 amino acids or >50% of its protein length aligned to the annotated transposons. For pseudogenes, we treated single-exon and multi-exon genes separately. For a single-exon gene, if its protein is fully covered by a multi-exon gene, it is designated as a pseudogene. For multi-exon gene, if its protein is fully covered by another gene, and its protein length is <70% of the latter, and its expression in RNA-seq data is lower than half of the average of all genes, it is designated as a pseudogene. For genes without protein evidence, a minimum cutoff of 50 amino acids was used to distinguish coding from non-coding genes. For single-exon genes with only EST or RNA-seq evidence support, we further filtered out those with protein length <100 amino acids and expression level in RNA-seq data lower than half of the average of all genes.

We categorized 37,516 genes (90.4%) with multiple types of evidence support as high-confidence and 3,991 genes (9.62%) with single type of evidence support as low-confidence. The number of genes supported by each evidence type is summarized in Extended Data Table 1a. Protein domains of each gene were annotated using InterProScan[57] by searching against publicly available databases, including ProDom (http://prodom.prabi.fr/), Prints (http://www.bioinf.manchester.ac.uk/dbbrowser/PRINTS/), Pfam (http://pfam.xfam.org/), Smart (http://smart.embl-heidelberg.de/), Panther (http://www.pantherdb.org/), Superfamily (http://supfam.org/SUPERFAMILY/), PIR (http://pir.georgetown.edu/) and Prosite (http://prosite.expasy.org/). Overall, 73.80% of the predicted proteins were found to contain InterPro domains. In addition, 55.08% of predicted genes have been classified by gene ontology terms and 11.35% of the genes were mapped to known biological pathways. Gene function was annotated according to the best matched proteins of *B. distachyon* (http://genome.jgi.doe.gov), and rice (http://rice.plantbiology.msu.edu/downloads_gad.shtml) using BLASTP with both minimum identity and coverage of 30% as thresholds (Extended Data Table 1c).

Segmentally and tandemly duplicated genes. To understand the chromosomal distribution of duplicated genes, we identified segmentally and tandemly duplicated genes in the Tu genome. Segmentally duplicated genes were identified in collinear segments, which contain at least two collinear gene pairs that were not separated by more than two non-collinear genes. Tandemly duplicated genes were paralogues that were located close to each other, and were not separated by more than two genes.

Orthologous genes between *T. urartu* and other grass genomes. We applied the standard OrthoMCL pipeline[58] to identify orthologous gene families among five grass species including *T. urartu*, rice, maize, sorghum and *B. distachyon*. The longest protein from each gene was selected, and the proteins with a length less than 30 amino acids were removed. After this step, pairwise sequence similarities between all input protein sequences were calculated using BLASTP with an E value cut-off of 1×10^{-5}. Markov clustering (MCL) of the resulting similarity matrix was used to define the orthologue cluster structure of the proteins, using an inflation value ($-I$) of 1.5 (OrthoMCL default). Then, comparative analysis was performed among *T. urartu*, rice, maize, sorghum and *B. distachyon* (Extended Data Fig. 4a)

Transcription factor analysis. To identify transcription factors in *T. urartu* genome, we blasted the annotated genes against known plant transcriptional factors collected from the iTAK database (http://itak.feilab.net/cgi-bin/itak/index.cgi). We then assigned these genes to specific transcription factor families using the prediction tool iTAK[14]. In total, 1,779 genes were classified as transcription factors into 68 families. To compare the size of transcriptional factor families among different species of grasses, we collected transcriptional factors in other six cereal genomes from iTAK (http://itak.feilab.net/cgi-bin/itak/index.cgi) including *B. distachyon*, *Oryza sativa*, *Sorghum bicolor*, *Zea mays*, *Aegilops tauschii* and *T. aestivum*, and investigated the enrichment of genes in each transcriptional factor family in all studied species (Supplementary Data 3).

Prolamin and disease resistance genes. The wheat prolamin gene sequences, including those encoding HMW-GS, LMW-GS, α-, γ-, ω- and δ-gliadin, were used as queries to blast against the *T. urartu* genome sequences with E value 1×10^{-10}, and matched sequences were extracted and manually annotated. Based on the annotation of whole gene set, we selected all of the NB-ARC domain genes and calculated their RPKM value based on the *T. urartu* RNA-seq data after inoculating with the powdery mildew pathogen *B. graminis* f. sp. *tritici*.

Gene expression profiling in leaf, root and spike of *T. urartu*. To study gene expression profiles, the RNA-seq data of leaves and roots of two-month-old plants and young spikes with 10–12 cm length were used[6]. Poor quality or technical sequences in Illumina paired-end reads (read length 75 bp) were removed using Trimmomatic version 0.35 preprocessing tool[59]. The qualified paired reads were then aligned against the IGDBv1.0 reference transcript sequences using Bowtie 2 version 2.2.6[60] to find all alignments of a read with no more than two mismatches. Subsequently, gene and isoform abundances were quantified from paired-end RNA-seq data using the RSEM software package[61]. Differentially expressed transcripts or genes with biological replicates were identified by running Bioconductor tools edgeR[62], which implements a range of statistical methods based on the negative binomial distributions. The differentially expressed genes were partitioned into clusters with dominantly high expression in one of the three tissues by perl script in Trinity[63].

Furthermore, we identified the functional preference of the dominantly expressed genes in each organ using a perl script included in Trinotate (http://trinotate.github.io/) to extract all Gene Ontology (GO) assignments from the TrEMBL/SwissProt databases, and used Bioconductor package GOseq to perform functional enrichment tests.

To monitor gene expression level along the *T. urartu* chromosomes in the three organs, we applied a window shift size of 5 Mb in a customized perl script. For each pseudomolecule, gene expression distribution among three organs is shown in Fig. 1k and Extended Data Fig. 3.

Comparison of *T. urartu* genome with BACs of *T. turgidum* and *T. aestivum*. To investigate sequence variation among different A genomes after polyploidization, we aligned the sequences of six BACs (881 kb in total) from the A subgenome of Tt and 11 BACs (1,423 kb in total) from the A subgenome of Ta to the Tu genome using NUCmer (parameter: -mum -mincluster 700) in MUMmer package[41]. Then we drew the dot plot using mummerplot in the same package with default parameters. We identified the repeat sequences using BLAST with E value 1×10^{-10} against the TREP database (http://botserv2.uzh.ch/kelldata/trep-db/index.html) and PGSB Repeat Element Database (http://pgsb.helmholtz-muenchen.de/plant/recat/) (Extended Data Fig. 5b).

We also compared the TGACv1 A subgenome sequences of hexaploid wheat[19] to the Tu genome. All scaffolds which were located on Ta7A were compared to Tu7. We performed all-to-all alignment (-minIdentity = 80~99, -minScore = 100, –fastMap) of Tu7 and Ta7A using BLAT[37]. The percentages of homologous segments on Tu7 and Ta7A were calculated using SOAP.COVERAGE[64] (Extended Data Fig. 5d).

Comparison of *T. urartu* genome with *T. aestivum* and *Ae. tauschii*. To investigate chromosomal structure variation between *T. urartu* and polyploid wheat, we compared the *T. urartu* genome with the three subgenomes (A, B, and D) of bread wheat (ftp://ftp.ensemblgenomes.org/pub/release-28/plants/fasta/triticum_aestivum/) as well with the D genome of *Ae. tauschii*[15]. Using software MCScanX[65] with at least three syntenic genes, we identified orthologous blocks and plotted homologous proteins among wheat genomes. Highly similar proteins (coverage of protein length ≥80 and identity ≥80) were obtained using BLAT[37] (Fig. 2a, b, Extended Data Fig. 5a).

Collinearity of *T. urartu* versus *B. distachyon, O. sativa* and *S. bicolor*. We identified homologous proteins between *T. urartu* and the other three genomes using BLASTP[66] with E value 1×10^{-5}, and scanned syntenic blocks consisted of homologous genes among the four genomes using MCScanX[65] with at least three syntenic genes (Extended Data Fig. 6).

Evolution of ancient duplicated blocks in *T. urartu*. We performed intragenomic comparison. Using all-against-all blastp and MCScanX[65] with default parameters in search of collinear paralogous relationships, five obvious collinear blocks were found in *T. urartu*. These blocks were then compared to seven previously published duplicated chromosome pairs in rice (Extended Data Fig. 7).

Comparisons of DNA and protein sequences between *T. urartu* chromosome 3 (Tu3) and *T. aestivum* chromosome 3B (Ta3B). Chromosome 3B of hexaploid wheat was completely sequenced and assembled using BAC-by-BAC sequencing strategy[26]. We performed complete and precise comparisons between Tu3 and Ta3B. The sequence of Ta3B was downloaded from the Ensembl website (ftp://ftp.ensemblgenomes.org/pub/release-28/plants/fasta/triticum_aestivum/). To compare the DNA sequence of the two chromosomes, we performed all-to-all alignment (-minIdentity = 80, -minScore = 100, –fastMap) of Tu3 and Ta3B with BLAT[37]. The percentages of homologous segments on Tu3 and Ta3B were calculated using SOAP.COVERAGE[64] (Extended Data Fig. 8a).

To compare the transposable element insertion date of Tu3 and Ta3B, we identified 5′ and 3′ solo-LTRs of Tu3 and Ta3B retrotransposons using LTRharvest[49]. The evolutionary distance of the two LTR sequences was estimated by using the Kimura two-parameter method[51]. A substitution rate of 1.3×10^{-8} mutations per site per year was used to convert evolutionary distance between 5′ and 3′ solo-LTRs to transposable element insertion dates[52] (Extended Data Fig. 8h).

Collinearity between Tu3 and Ta3B. We plotted homologous DNA sequences between Tu3 and Ta3B using NUCmer (parameter: -mum -mincluster 700) in MUMmer package[41]. Collinearity in both regions of 0–200-Mb and 400–700-Mb segments can be clearly observed between two chromosomes. Within 200–400-Mb segments, collinearity was not identified between Tu3 and Ta3B.

We also explored collinear relationships of homologous genes between Tu3 and Ta3B. Syntenic blocks containing at least three homologous gene pairs were identified using MCScanX[65] with at least three syntenic genes (Extended Data Fig. 8b–g).

Identification of gene insertions and deletions on Tu3 and Ta3B. We computationally identified gene insertions and deletions on Tu3 relative to Ta3B by combining data from *B. distachyon* (Bdistachyon_283_v2.1), rice (IRGSP1.0) and sorghum (http://phytozome.jgi.doe.gov/pz/portal.html). They are grass relatives of wheat that have well-sequenced and annotated genomic data.

MCScanX[65] was used to identify collinear duplicated gene blocks between species or genomes, including Tu3-to-Ta3B, Tu3-to-Bd, Tu3-to-Os, Tu3-to-Sb, Ta3B-to-Bd, Ta3B-to-Os and Ta3B-to-Sb. We collected 176 syntenic blocks between Tu3 and Ta3B, each containing more than five paired orthologues. Customized perl scripts were used to identify gene insertions and deletions occurred in the syntenic blocks between Tu3 and Ta3B via detecting whether the orthologues in Bd, Os and Sb exist. If a Ta3B gene had orthologues in at least two of Bd, Os and Sb but not on Tu3, a gene deletion was defined to have occurred on Tu3. A Tu3 insertion was defined if no collinear orthologues were found in any other investigated genomes, and the orthologues of two adjacent genes around the inserted segment existed in a collinear region of at least two of Bd, Os and Sb.

Analysis of *T. urartu* populations. A total of 147 *T. urartu* accessions, collected from Armenia, Iran, Iraq, Syria, Turkey and Lebanon, were used in this study (Supplementary Data 8). Leaf samples from five uniform seedlings were used for total RNA extraction with Illumina TruSeq RNA Sample Prep Kit. With them, 147 paired-end libraries were constructed using Illumina Paired-End Sample Prep Kit, and sequenced by Illumina HiSeq 2000 platform. In total, 63 billion paired-end reads in length of 100 bp were generated (6.3 Tb of sequences), with an average coverage depth of more than 25× for each accession. Then, adaptor sequence trimming and removal of low-quality reads were performed with the ngsShoRT algorithms[67].

After removing adaptor sequences and reads with low sequence quality, TopHat2[68] was used to map the paired-end reads against the reference sequence of G1812. Only paired-end reads that mapped uniquely to the genome were used for further analysis of variation calling. Duplicated reads were also filtered. The SNP calling were performed by SAMtools mpileup package[69], and SNPs with minor allele frequency lower than 5% were excluded from further analyses.

For population structure analysis, the neighbour-joining tree was constructed using MEGA5 based on all of the SNPs[70]. Population structure was calculated using STRUCTURE software[71]. The number of genetic clusters K was predefined as 1–10 to explore the population structure with three iterations. The run with maximum likelihood was used to assign individual genotypes into groups. The three groups uncovered using STRUCTURE corresponded well to those based on phylogenetic clustering with respect to accession composition in each group. The statistics of sequence diversity (F_{ST}) and the population differentiation (π and θ) were computed using a 100-kb window in 10-kb steps with the PopGen package in BioPerl[72].

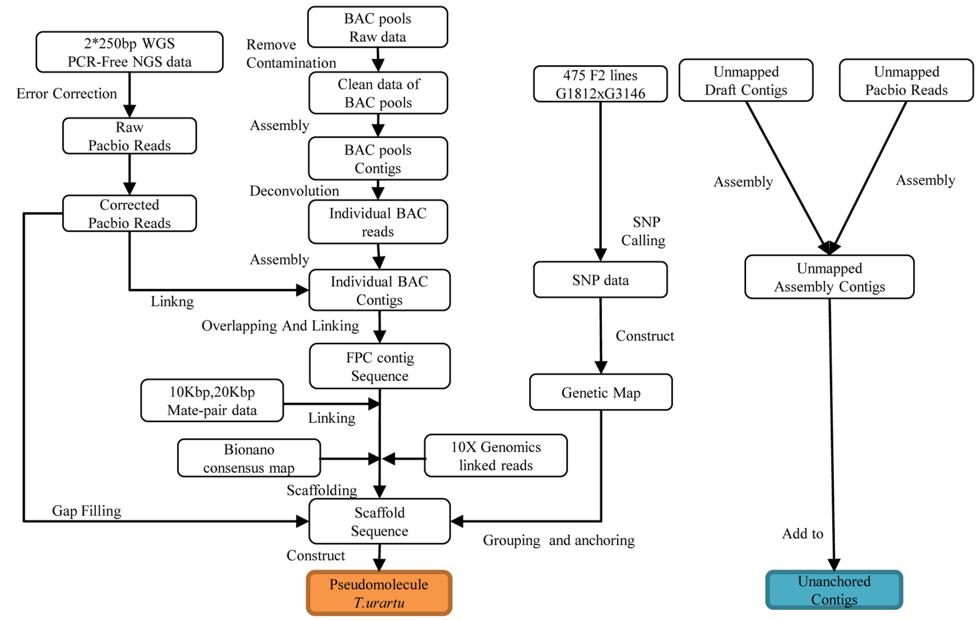
Reporting summary. Further information on experimental design is available in the Nature Research Reporting Summary linked to this paper.

Data availability. Sequence data and assemblies have been deposited at BioProject under project accession number PRJNA337888, Sequence Read Archive SRP081049 (PacBio reads SRR4010673–SRR4010781 and Illumina PCR-free reads SRR4010671–SRR4010672), and GenBank MKGO00000000 (pseudomolecules of *T. urartu*). The Tu genome annotation is available at MBKBase website (http://www.mbkbase.org/Tu/). BAC assemblies and the sequence reads of the population (including SNPs) have been deposited at GSA (http://gsa.big.ac.cn/) with the accession number PRJCA000369.

31. Akhunov, E. D., Akhunova, A. R. & Dvorák, J. BAC libraries of *Triticum urartu, Aegilops speltoides* and *Ae. tauschii*, the diploid ancestors of polyploid wheat. *Theor. Appl. Genet.* **111**, 1617–1622 (2005).
32. Zhang, H. B., Zhao, X. P., Ding, X. L., Paterson, A. H. & Wing, R. A. Preparation of megabase-size DNA from plant nuclei. *Plant J.* **7**, 175–184 (1995).
33. van Oeveren, J. et al. Sequence-based physical mapping of complex genomes by whole genome profiling. *Genome Res.* **21**, 618–625 (2011).
34. Soderlund, C., Longden, I. & Mott, R. FPC: a system for building contigs from restriction fingerprinted clones. *Comput. Appl. Biosci.* **13**, 523–535 (1997).
35. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
36. Zimin, A. V. et al. The MaSuRCA genome assembler. *Bioinformatics* **29**, 2669–2677 (2013).
37. Kent, W. J. BLAT–the BLAST-like alignment tool. *Genome Res.* **12**, 656–664 (2002).
38. Hackl, T., Hedrich, R., Schultz, J. & Förster, F. proovread: large-scale high-accuracy PacBio correction through iterative short read consensus. *Bioinformatics* **30**, 3004–3011 (2014).
39. Chaisson, M. J. & Tesler, G. Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory. *BMC Bioinformatics* **13**, 238 (2012).
40. Zimin, A. V. et al. Hybrid assembly of the large and highly repetitive genome of *Aegilops tauschii*, a progenitor of bread wheat, with the MaSuRCA mega-reads algorithm. *Genome Res.* **27**, 787–792 (2017).
41. Delcher, A. L., Phillippy, A., Carlton, J. & Salzberg, S. L. Fast algorithms for large-scale genome alignment and comparison. *Nucleic Acids Res.* **30**, 2478–2483 (2002).
42. Boetzer, M., Henkel, C. V., Jansen, H. J., Butler, D. & Pirovano, W. Scaffolding pre-assembled contigs using SSPACE. *Bioinformatics* **27**, 578–579 (2011).
43. English, A. C. et al. Mind the gap: upgrading genomes with Pacific Biosciences RS long-read sequencing technology. *PLoS One* **7**, e47768 (2012).
44. McKenna, A. et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
45. Stam, P. Construction of integrated genetic linkage maps by means of a new computer package: Join Map. *Plant J.* **3**, 739–744 (1993).
46. Wu, Y., Bhat, P. R., Close, T. J. & Lonardi, S. Efficient and accurate construction of genetic linkage maps from the minimum spanning tree of a graph. *PLoS Genet.* **4**, e1000212 (2008).
47. Xu, Z. & Wang, H. LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Res.* **35**, W265–W268 (2007).

48. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.* **27**, 573–580 (1999).
49. Ellinghaus, D., Kurtz, S. & Willhoeft, U. LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. *BMC Bioinformatics* **9**, 18 (2008).
50. Chenna, R. et al. Multiple sequence alignment with the Clustal series of programs. *Nucleic Acids Res.* **31**, 3497–3500 (2003).
51. Yang, Z. PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput. Appl. Biosci.* **13**, 555–556 (1997).
52. Ma, J. & Bennetzen, J. L. Rapid recent growth and divergence of rice nuclear genomes. *Proc. Natl Acad. Sci. USA* **101**, 12404–12410 (2004).
53. Lowe, T. M. & Eddy, S. R. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.* **25**, 955–964 (1997).
54. Nawrocki, E. P., Kolbe, D. L. & Eddy, S. R. Infernal 1.0: inference of RNA alignments. *Bioinformatics* **25**, 1335–1337 (2009).
55. Kong, L. et al. CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine. *Nucleic Acids Res.* **35**, W345–W349 (2007).
56. Trapnell, C. et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* **28**, 511–515, (2010).
57. Jones, P. et al. InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**, 1236–1240, (2014).
58. Li, L., Stoeckert, C. J., Jr & Roos, D. S. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res.* **13**, 2178–2189 (2003).
59. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
60. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10**, R25 (2009).
61. Li, B. & Dewey, C. N. RSEM: accurate transcript quantification from RNA-seq data with or without a reference genome. *BMC Bioinformatics* **12**, 323 (2011).
62. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2010).
63. Grabherr, M. G. et al. Full-length transcriptome assembly from RNA-seq data without a reference genome. *Nat. Biotechnol.* **29**, 644–652 (2011).
64. Luo, R. et al. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *Gigascience* **1**, 18 (2012).
65. Wang, Y. et al. MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Res.* **40**, e49 (2012).
66. Altschul, S. F. et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402 (1997).
67. Chen, C., Khaleel, S. S., Huang, H. & Wu, C. H. Software for pre-processing Illumina next-generation sequencing short read sequences. *Source Code Biol. Med.* **9**, 8 (2014).
68. Kim, D. et al. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.* **14**, R36 (2013).
69. Li, H. et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
70. Tamura, K. et al. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol. Biol. Evol.* **28**, 2731–2739 (2011).
71. Pritchard, J. K., Stephens, M. & Donnelly, P. Inference of population structure using multilocus genotype data. *Genetics* **155**, 945–959 (2000).
72. Wright, S. The genetical structure of populations. *Ann. Eugen.* **15**, 323–354 (1951).

a



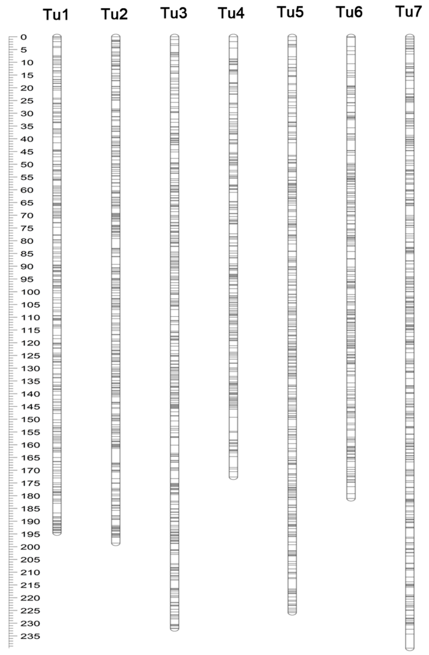
b

Data type	LIS	ARL (bp)	RD (Gb)	UD (Gb)	ED*
BAC	~ 300bp	150	2,102	1,471	297
WGS	~ 500bp	250	130	107	21
PacBio	----	8,120		97	19
BioNano	----	294,945		417	83
10x Genomics	----	31,392		57	11
Total	----	----	2,232	2,149	431

c

	Contig*		Scaffold#	
	Size (bp)	Number	Size (bp)	Number
N90	66,141	15,319	85,682	3,045
N80	136,141	10,607	1,485,732	995
N70	199,189	7,702	2,251,206	736
N60	265,486	5,620	2,959,487	552
N50	344,138	4,032	3,670,573	405
Longest	3,002,449	----	18,759,549	----
Total number	----	40,796	----	31,559
Total size	4,793,303,795	----	4,860,907,502	----
Size on Chr	4,633,757,067	----	4,665,678,060	----
GC ratio (%)	45.93	----	45.93	----

d



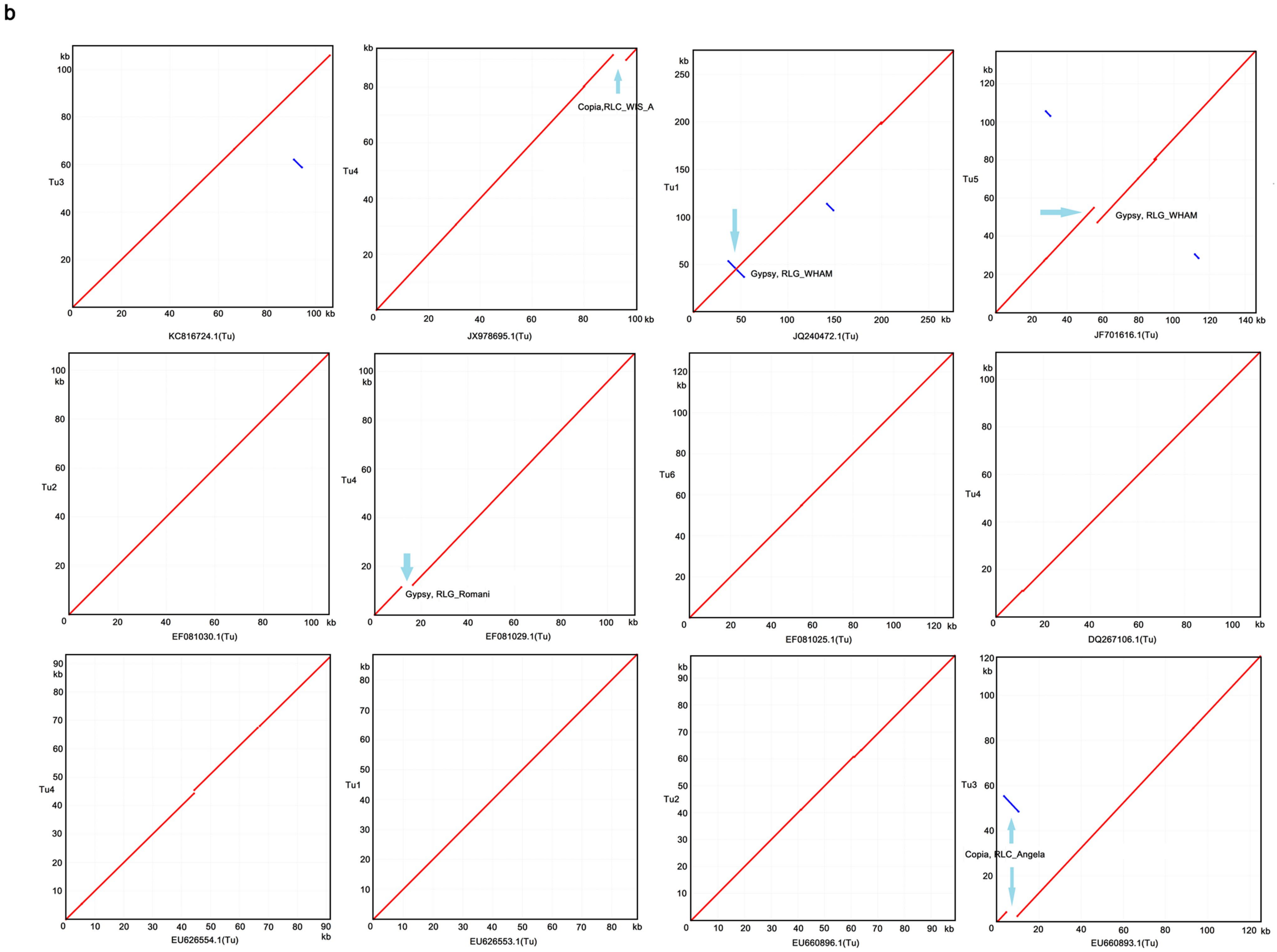
e

Chr	Chr length (bp)	Bin#	GLDAccu (cM)	GLDAvg (cM)	PDPCM	Gene#
Tu1	583,994,449	404	194.58	0.30	3,001,369	4,885
Tu2	755,818,500	937	198.60	0.14	3,795,643	6,396
Tu3	747,043,632	695	232.14	0.21	3,218,018	5,931
Tu4	619,528,116	520	172.69	0.22	3,587,599	4,036
Tu5	662,474,153	680	225.82	0.20	2,929,158	5,720
Tu6	576,970,508	503	181.10	0.23	3,180,471	4,467
Tu7	719,848,703	767	239.78	0.19	3,002,097	6,141
Total	4,665,678,060	4,506	1,444.71	0.21	3,226,725	37,576

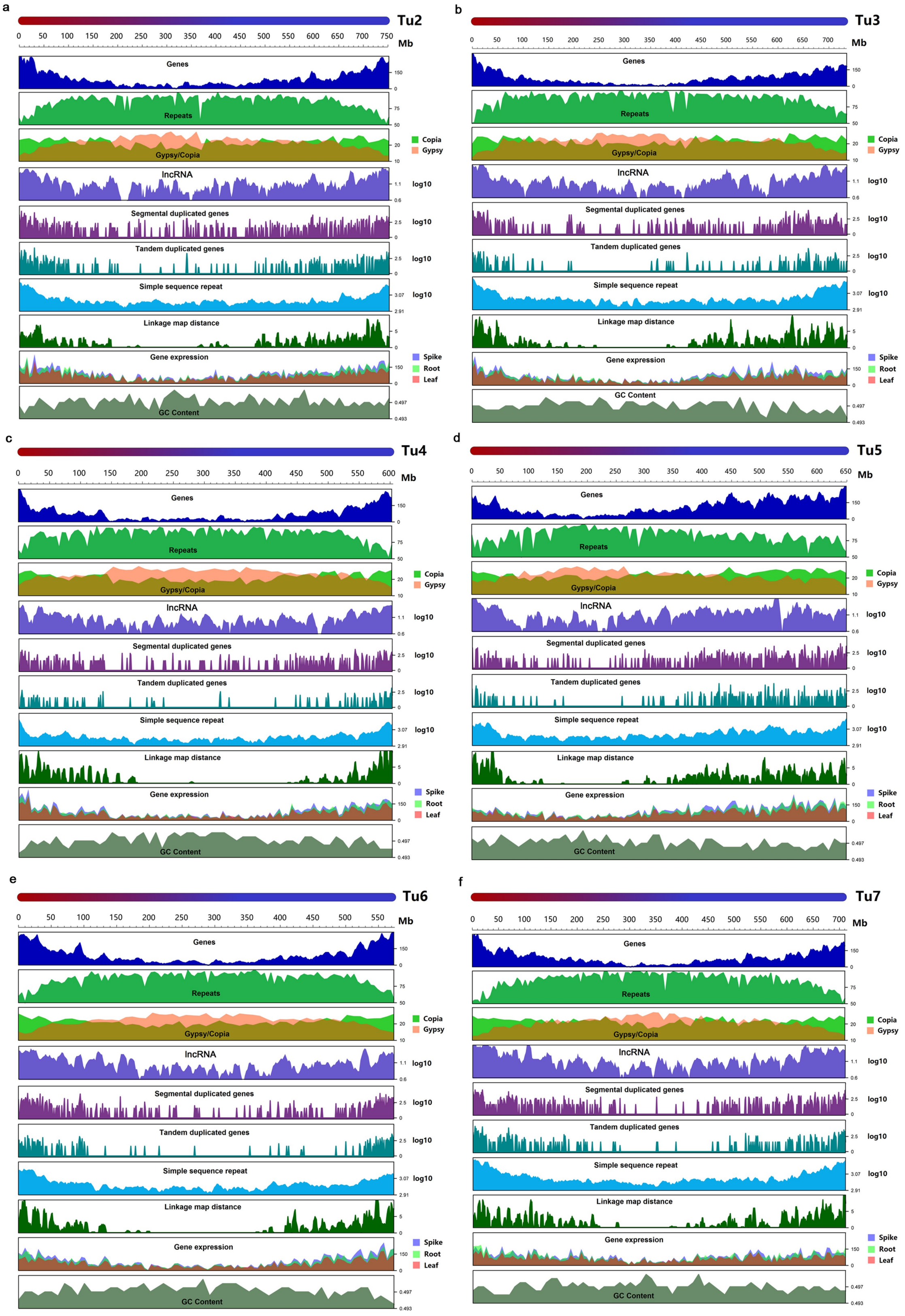
Extended Data Fig. 1 | *T. urartu* genome assembly. a, Schematic workflow for genome sequencing, assembly and chromosomal assignment with a high-density SNP map. **b**, Statistics of sequencing data. *Calculated from the estimated genome size of 4.94 Gb. LIS, library insert size; ARL, average read length; RD, raw data; UD, usable data; ED, effective depth. **c**, Summary of the Tu genome assembly. *Contig: contiguous sequence without Ns assembled with Illumina reads, corrected by PacBio reads. #Scaffold: Sequence with Ns, in which two or more contigs were connected by mate-pair reads, BioNano genome maps and 10× Genomics linked reads. **d**, High-resolution genetic map of *T. urartu* using SNP markers. The SNP markers were identified from an F2 population (475 individuals) derived from a cross between accessions G1812 and G3146 of Tu. **e**, Summary of physical length and genetic map of seven pseudomolecules. Chr, chromosome; GLDAccu, genetic linkage distance accumulation; GLDAvg, genetic linkage distance on average; PDPCM, physical distance per centiMorgen (cM).

a

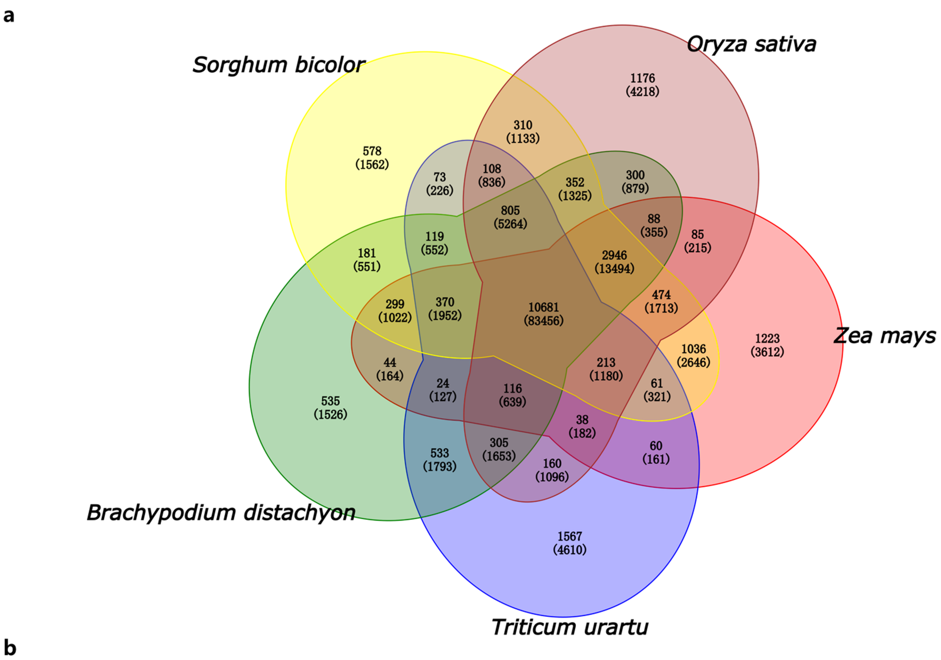
Species	GenBank accession	Chr	Start (bp)	End (bp)	BAC length (bp)	Coverage (bp)	Coverage (%)	Identity (%)	Coverage draft# assembly (bp)	Coverage draft# assembly (%)	Identity (%)	Chr*
Tu	KC816724.1	Tu3	699,456,034	699,562,017	107,101	105,947	98.92	99.34	68,467	63.93	99.80	Tu3L
Tu	JX978695.1	Tu4	21,129,670	21,223,255	100,141	95,060	94.93	99.85	68,048	67.95	99.69	NA
Tu	JQ240472.1	Tu1	510,256,350	510,530,722	275,997	275,532	99.83	99.67	142,217	51.53	99.32	NA
Tu	JF701616.1	Tu5	604,310,067	604,446,989	145,644	143,703	98.66	99.24	85,592	58.77	99.40	NA
Tu	EF081030.1	Tu2	643,521,646	643,628,384	106,806	106,741	99.94	99.98	64,999	60.86	99.79	Tu2
Tu	EF081029.1	Tu4	251,386,647	251,493,726	111,168	106,222	95.55	99.23	93,779	84.36	99.53	Tu4
Tu	EF081025.1	Tu6	353,626,321	353,755,476	129,021	129,019	100	99.91	83,004	64.33	99.38	Tu6
Tu	DQ267106.1	Tu4	415,179,777	415,291,155	111,912	111,425	99.56	99.60	62,696	56.02	99.77	Tu4
Tu	EU626554.1	Tu4	593,226,325	593,318,340	91,075	90,089	98.91	99.93	29,170	32.03	99.85	NA
Tu	EU626553.1	Tu1	554,496,112	554,584,624	88,460	88,459	100	99.83	80,277	90.75	99.79	NA
Tu	EU660896.1	Tu2	28,849,398	28,947,703	98,890	98,277	99.38	99.86	73,664	74.49	99.37	Tu2
Tu	EU660893.1	Tu3	703,126,808	703,244,142	124,885	124,195	99.45	99.53	76,065	60.91	99.46	Tu3L



Extended Data Fig. 2 | Evaluation of *T. urartu* genome assembly. **a**, Summary of comparison of the *T. urartu* genome assembly with public BAC sequences. *The published chromosome location of BACs; NA, not available. #Tu draft assembly[6]. **b**, Dot plots showing comparison of *T. urartu* genome with available BACs of *T. urartu* from public database. The blue arrows indicate the regions at which the BAC sequences and pseudomolecules did not match owing to the presence of repeat elements.



Extended Data Fig. 3 | Chromosomal distribution of *T. urartu* genome features. a–f, Features on Tu2–Tu7 are in the order of DNA pseudomolecule; gene frequency (number of genes per 10 Mb); repeat density (per cent nucleotides per 5 Mb); density of LTR retrotransposons (per cent nucleotides per 10 Mb); frequency of lncRNA (log[number of genes per 10 Mb]); frequency of segmentally duplicated genes (log[number of genes per 1 Mb]); frequency of tandemly duplicated genes (log[number of genes per 1 Mb]); frequency of simple sequence repeats (log[number of repeats per 10 Mb]); linkage map distance (cM per 5 Mb); accumulated gene expression level in RNA-seq data ($\log_2$[FPKM per 5 Mb]); GC content (per cent per 1 Mb).



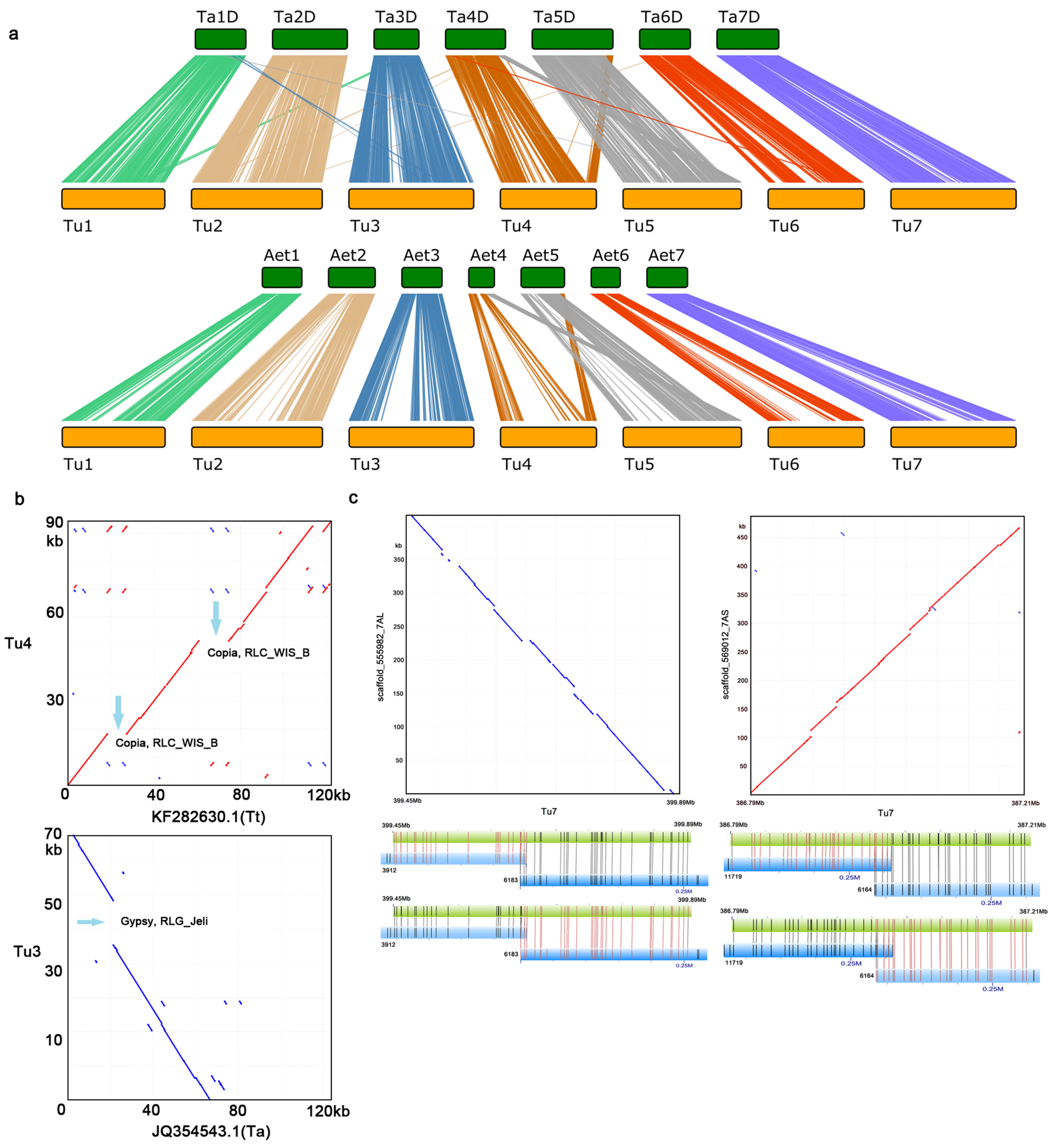
b

category	over represented pvalue	Tu-specific InCat#	Total InCat#	term
GO:0006952	2.40E-22	492	2125	defense response
GO:0006955	1.15E-11	137	599	immune response
GO:0008219	3.00E-10	124	540	cell death
GO:0050896	4.70E-10	1032	5293	response to stimulus
GO:0006950	1.03E-08	798	4039	response to stress
GO:0043412	5.23E-08	457	2280	macromolecule modification
GO:0005524	3.51E-50	994	4504	ATP binding
GO:0097367	4.31E-45	1057	4885	carbohydrate derivative binding
GO:0032553	6.19E-45	1040	4821	ribonucleotide binding
GO:0000166	2.97E-43	1140	5424	nucleotide binding
GO:0004672	1.10E-33	487	2023	protein kinase activity
GO:0016773	4.23E-33	517	2186	phosphotransferase activity, alcohol group as acceptor
GO:0043531	4.02E-27	199	772	ADP binding
GO:0043167	3.74E-15	1896	9884	ion binding
GO:0016705	3.89E-10	122	598	oxidoreductase activity, acting on paired donors
GO:0004497	7.47E-10	108	530	monooxygenase activity
GO:0097159	5.09E-09	1862	9857	organic cyclic compound binding
GO:1901363	5.53E-09	1861	9852	heterocyclic compound binding
GO:0016740	9.78E-09	1044	5527	transferase activity

c

B3_subfamily	Bd	Os	Sb	Zm	Tu	Aet	Ta
ARF	0	6	0	0	0	0	3
RAV	4	12	10	16	9	11	13
REM	14	24	39	16	61	53	45
LAV	6	12	9	11	8	4	6
Others*	25	0	0	11	29	28	71
Total	49	54	58	54	107	96	138

Extended Data Fig. 4 | Analyses of gene families and B3 transcription factors. a, Comparison of gene families of *T. urartu* with *O. sativa*, *Z. mays*, *S. bicolor* and *B. distachyon*. Venn diagram illustrates shared and unique gene families (gene numbers in parentheses) among the five grass species. **b**, Gene ontology analysis of Tu-specific genes. Tu-specific InCat#, number of Tu-specific genes in the GO category; Total InCat#, number of total Tu genes in the GO category. **c**, Comparison of B3 transcription factors of Tu with *B. distachyon* (Bd), *O. sativa* (Os), *S. bicolor* (Sb), *Z. mays* (Zm), *Ae. tauschii* (Aet) and *T. aestivum* (Ta). *B3 transcription factors without identified subfamily.

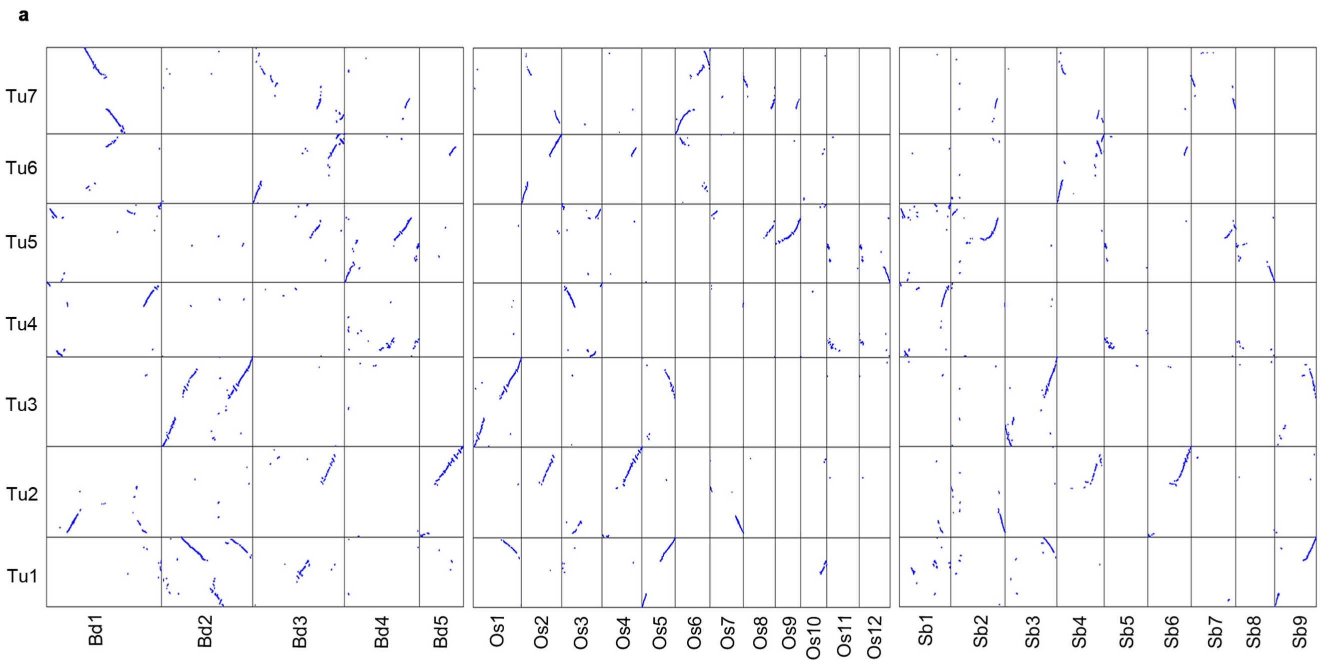


d

Nucleotide[*]		Tu7			TGACv1 Ta7A Scaffolds		
Identity (%)	ML[†]	Coverage (bp)	Total length (excluded N)	Coverage (%)	Coverage (bp)	Total length (excluded N)	Coverage (%)
80	100	658,438,880	719,689,837	91.49	538,525,255	594,803,487	90.54
85	100	658,295,052	719,689,837	91.47	538,387,524	594,803,487	90.52
90	100	655,109,129	719,689,837	91.03	535,706,888	594,803,487	90.06
93	100	641,746,482	719,689,837	89.17	526,338,048	594,803,487	88.49
95	100	610,625,025	719,689,837	84.85	506,102,676	594,803,487	85.09
96	100	565,997,619	719,689,837	78.64	476,353,402	594,803,487	80.09
97	100	443,480,816	719,689,837	61.62	385,367,214	594,803,487	64.79
98	100	269,892,779	719,689,837	37.50	244,672,441	594,803,487	41.14
99	100	144,732,287	719,689,837	20.11	130,521,272	594,803,487	21.94
100	100	84,814,119	719,689,837	11.78	74,258,058	594,803,487	12.48

Extended Data Fig. 5 | See next page for caption.

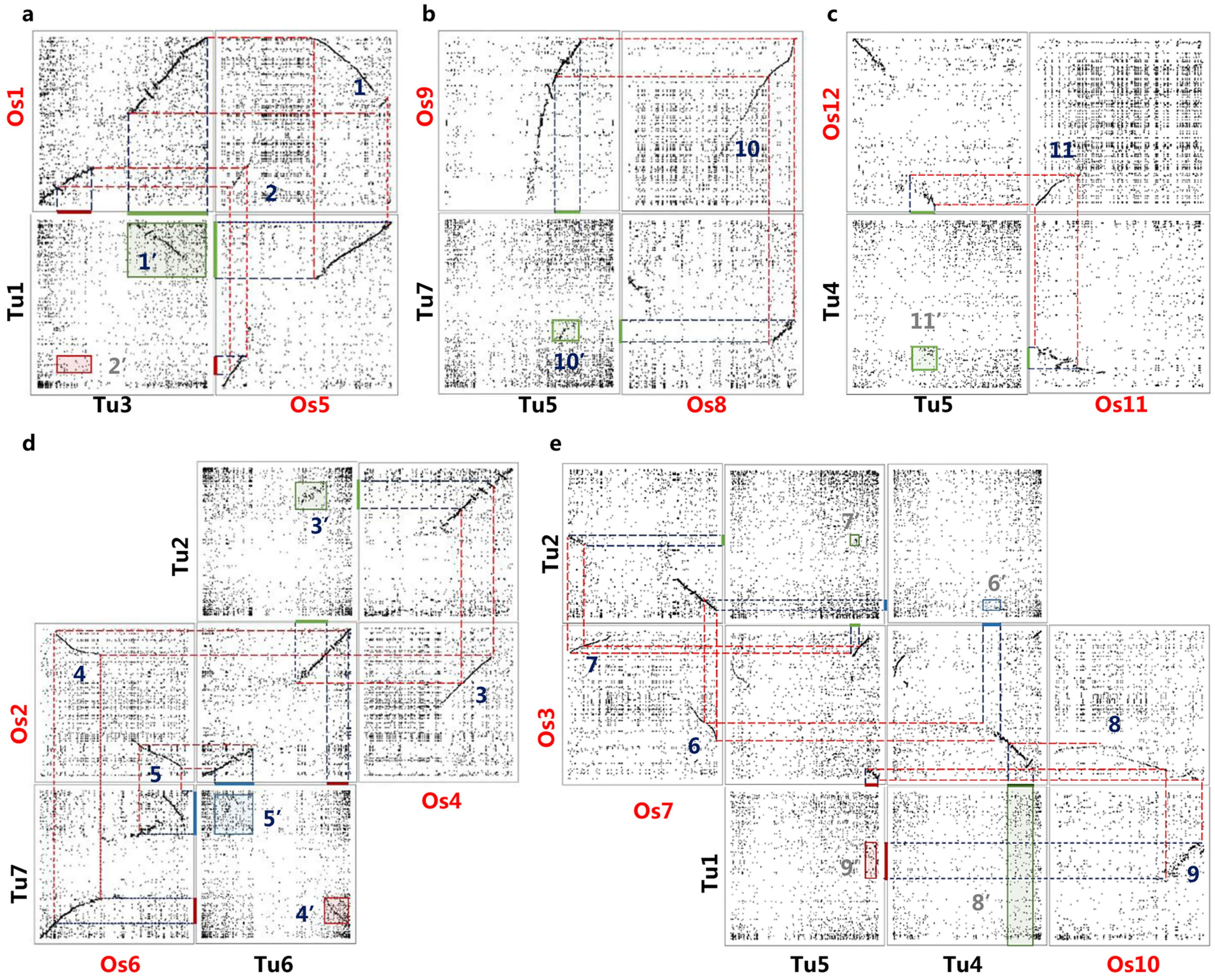
Extended Data Fig. 5 | Comparison of Tu genome with other wheat genomes. a, Syntenic analysis of Tu genome with the D subgenome of Ta, and the genome of Aet. Each syntenic block contains five or more genes, with sequence similarity of 80% or more. **b**, Comparison of Tu genome with BACs of *T. turgidum* (Tt) and Ta. BAC KF282630 from chromosome 4 of the A subgenome of Tt contained two inserted fragments (blue arrows) that are composed of Copia RLC_WIS_B elements, which were not detected on the corresponding Tu4 region. BAC JQ354543 from chromosome 3 of the A subgenome of Ta lacked the Gypsy RLG_Jeli element (blue arrow), which was found on the corresponding region of Tu3. **c**, Comparison of the Tu genome with Ta7A scaffolds from TGACv1[19]. The dot plots on the top show comparison of two largest TaA scaffolds to corresponding parts of Tu7 chromosome. The diagonal lines on the dot plots show fine co-linearity. The lower part shows validation of the sequence assembly of Tu7 by BioNano maps. The Tu7 sequences were digested into in silico consensus maps, and the consensus maps corresponding to the two Ta7A scaffolds (green bar) are compared against their corresponding BioNano genome maps (blue bar). Each vertical line on the green/blue bars indicates a restriction enzyme cutting site (Nt.BspQI), and vertical lines between green bars and blue bars indicate alignments among these sites. The blue bars highlighted with red vertical line demonstrate that two BioNano genome maps (for example, 3912 and 6183) overlap with one another (they all have alignments on the overlapping region), although the two maps are not merged together owing to lack of coverage on the overlapping region. The high consistency of alignments between consensus maps and BioNano genome maps confirm the high quality of Tu genome assembly. Therefore, the insertion/deletion events in the dot plots should be sequence variations between the two A genomes from Ta and Tu, rather than assembly errors. **d**, Comparison of Tu7 with all Ta7A scaffolds from TGACv1 at nucleotide levels. *Nucleotide: minimum cutoff of DNA sequence alignments between Tu7 and Ta7A. †ML, minimum length (bp) to align.



b

Syntenic block	Tu		Bd		Os		Sb	
	Chr	Range (Mb)	Chr	Range (Mb)	Chr	Range (Mb)	Chr	Range (Mb)
1	Tu1	0-220	Bd2	31-40	Os5	0-8	Sb9	0-12
2	Tu1	116-374	Bd3	27-36	Os10	14-23	Sb1	16-53
3	Tu1	373-569	Bd2	12-29	Os5	16-29	Sb9	40-59
4	Tu2	0-21	Bd5	0-7	Os4	0-10	Sb6	0-6
5	Tu2	20-374	Bd1	13-22	Os7	18-29	Sb2	66-77
6	Tu2	347-420	Bd1	4-6	Os7	0-4	Sb2	0-6
7	Tu2	422-740	Bd5	7-28	Os4	16-35	Sb6	36-62
8	Tu3	1-247	Bd2	0-12	Os1	0-10	Sb3	0-12
9	Tu3	239-733	Bd2	40-59	Os1	20-43	Sb3	48-74
10	Tu4	5-88	Bd1	4-10	Os3	25-31	Sb1	5-14
11	Tu4	51-141	Bd4	22-31	Os11	0-11	Sb5	0-12
12	Tu4	134-591	Bd1	60-72	Os3	2-16	Sb1	56-72
13	Tu4	565-604	Bd1	0-3	Os3	34-36	Sb1	0-3
14	Tu5	0-160	Bd4	0-6	Os12	17-27	Sb8	41-55
15	Tu5	165-317	Bd4	44-48	Os12	0-4	Sb8	0-10
16	Tu5	315-530	Bd4	32-43	Os9	0-22	Sb2	44-67
17	Tu5	533-606	Bd1	2-6	Os3	30-34	Sb1	2-7
18	Tu5	605-648	Bd1	72-74	Os3	0-2	Sb1	70-73
19	Tu6	1-375	Bd3	0-8	Os2	0-9	Sb4	0-13
20	Tu6	195-570	Bd3	47-59	Os2	20-35	Sb4	49-68
21	Tu7	0-205	Bd1	38-50	Os6	0-13	Sb10	0-24
22	Tu7	168-329	Bd3	41-44	Os8	24-28	Sb7	56-64
23	Tu7	315-470	Bd3	11-15	Os8	0-5	Sb7	0-8
24	Tu7	459-709	Bd1	24-36	Os6	14-31	Sb10	40-60

Extended Data Fig. 6 | Synteny of *T. urartu* genome with other grass genomes. a, Intergenomic dot plots showing orthologous syntenies between Tu and three grass relatives *B. distachyon* (Bd), *O. sativa* (Os) and *S. bicolor* (Sb). The alignment of the Tu genome to the Bd, Os and Sb genomes demonstrates their highly collinear relationships and full coverage of Tu by orthologous chromosome segments from each of the three grass genomes. **b**, Data table showing the orthologous syntenies between Tu and the other genomes (Bd, Os and Sb).



f

Ancient chromosomes*	Tu regions		Os regions	
	Chr	Location (Mb)	Chr	Location (Mb)
A1/A5	**Tu1**	**434-557** 106-224	Os5	19-29 3-8**
	Tu3	**389-634** 50-118	Os1	20-38 2-7**
A2/A4	**Tu2**	**548-670**	Os4	18-30
	Tu6	**221-466**	Os2	17-28
A2/A6	**Tu6**	**2-141 471-548**	Os2	0-9 29-35
	Tu7	**578-680 78-172**	Os6	17-31 3-12
A3/A7	Tu2	33-134	Os7	23-29**
	Tu4	55-462	Os3	9-19**
	Tu2	208-396	Os7	0-11**
	Tu5	548-588	Os3	31-36**
A3/A10	Tu1	92-374	Os10	7-23**
	Tu4	410-573	Os3	0-6**
	Tu1	94-372	Os10	23-27**
	Tu5	523-603	Os3	22-29**
A8/A9	**Tu5**	**346-523**	Os9	3-22
	Tu7	**201-282**	Os8	13-28
A11/A12	Tu4	66-143	Os11	0-12**
	Tu5	142-339	Os12	0-7**

g

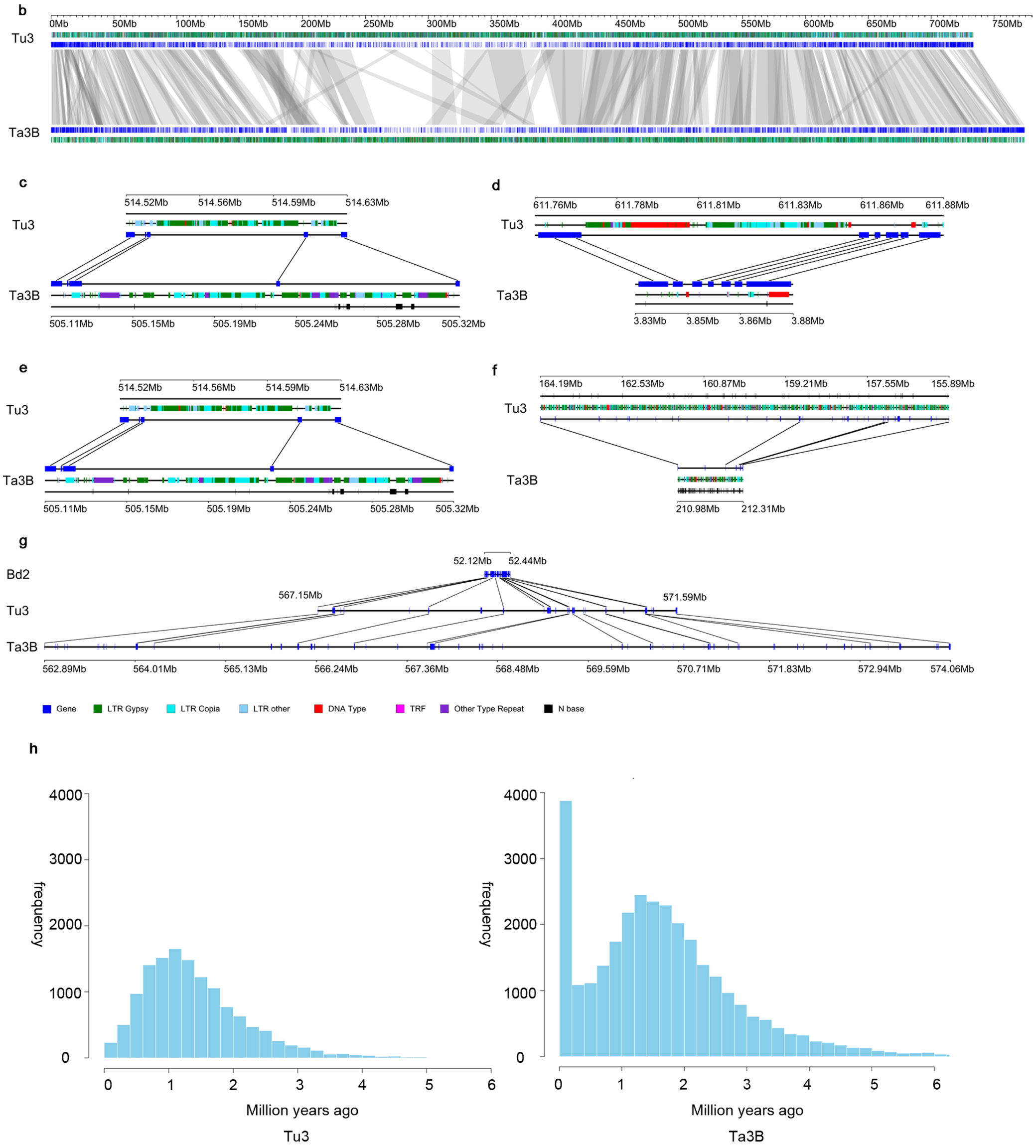
Ancient chromosomes*	Os syntenic gene pairs	Os Chr	Tu Chr	Tu orthologous genes (%)	Tu syntenic gene pairs (%)	Single copy loss (%)	Double copy loss (%)
A1/A5	693	Os1	Tu3	310 (45)	147 (21)	336 (49)	210 (30)
		Os5	Tu1	320 (46)			
A2/A4	378	Os2	Tu6	107 (28)	60 (16)	193 (51)	125 (33)
		Os4	Tu2	206 (54)			
A2/A6	425	Os2	Tu6	171 (40)	95 (22)	194 (46)	136 (32)
		Os6	Tu7	213 (50)			
A3/A7	301	Os3	Tu4/5	111 (37)	37 (12)	166 (55)	98 (33)
		Os7	Tu2	129 (43)			
A3/A10	182	Os3	Tu4/5	73 (40)	20 (11)	82 (45)	80 (44)
		Os10	Tu1	49 (27)			
A8/A9	305	Os8	Tu7	62 (20)	29 (10)	145 (47)	131 (43)
		Os9	Tu5	141 (46)			
A11/A12	336	Os11	Tu4	31 (9)	1 (0)	124 (37)	211 (63)
		Os12	Tu5	95 (28)			
Total	2620				389 (15)	1240 (47)	991 (38)

Extended Data Fig. 7 | Intragenomic collinear regions of Tu corresponding to rice (Os) duplications. a, Intragenomic and intergenomic dot plots show a clearly visible collinear region 1′ between Tu1 and Tu3 that is orthologous to the Os intragenomic collinear region 1. However, the collinearity of region 2′ between Tu1 and Tu3 is severely corrupted, but the corresponding Os intragenomic collinear region 2 was clearly visible. **b**–**e**, Similarly, Tu intragenomic collinear regions 10′, 3′, 4′ and 5′ were also clearly visible, which correspond to Os intragenomic collinear regions 10, 3, 4 and 5, respectively. However, the collinearity of Tu intragenomic regions corresponding to Os intragenomic collinear regions 11, 6, 7, 8 and 9 was disrupted. **f**, Data table showing the strong intragenomic collinear segments of Os and their corresponding regions of Tu (chromosomal positions are shown in Mb). *The corresponding seven pairs of ancient chromosomes based on the AGK structure[24] that are ancestral to the Tu and Os chromosomes listed in the right columns. The five clearly visible regions in Tu are marked in bold. **Os regions without visible corresponding collinear regions in Tu dot plots. **g**, Data table showing the number of collinear genes between ancestrally duplicated chromosome segments in rice and their corresponding gene numbers in Tu.

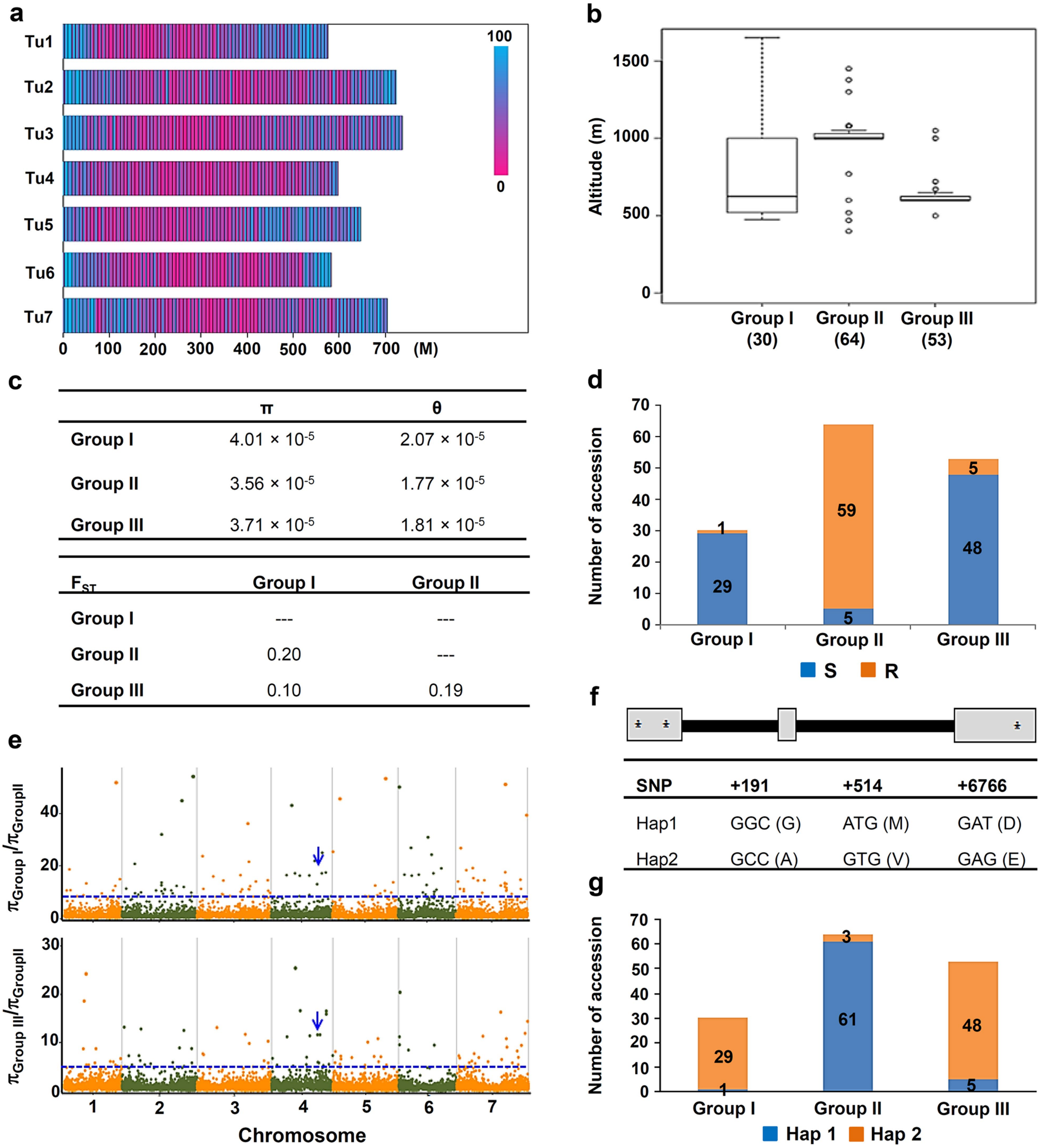
a

Nucleotide*		Tu3			Ta3B		
Identity (%)	ML†	Coverage (bp)	Chr length (exclued N)	Coverage (%)	Coverage (bp)	Chr length (excluded N)	Coverage (%)
80	100	574,638,057	742,462,632	77.40	517,342,230	723,068,628	71.55
90	100	537,678,548	742,462,632	72.42	463,022,642	723,068,628	64.04
80	500	549,749,814	742,462,632	74.04	484,964,143	723,068,628	67.07
90	500	502,467,297	742,462,632	67.68	417,218,269	723,068,628	57.70
80	1000	487,349,674	742,462,632	65.64	414,524,058	723,068,628	57.33
90	1000	435,660,075	742,462,632	58.68	346,724,228	723,068,628	47.95

Protein‡		Tu3			Ta3B		
Identity (%)	Coverage (%)	Aligned gene#	Gene#	%	Aligned gene#	Gene#	%
25	25	3,602	5,931	60.73	4,158	6,684	62.21
35	35	3,449	5,931	58.15	3,935	6,684	58.87
50	50	3,103	5,931	52.32	3,542	6,684	52.99
90	90	1,721	5,931	29.02	2,017	6,684	30.18



Extended Data Fig. 8 | Comparison of chromosome 3 of Tu with chromosome 3B of Ta. a, Comparison of Tu3 with Ta3B at both nucleotide and protein levels. *Nucleotide: minimum cutoff of DNA sequence alignments between Tu3 and Ta3B. †ML, minimum length (bp) to align. ‡Protein: minimum cutoff of protein sequence alignments between Tu3 and Ta3B, and in reverse. **b**, Overall view of syntenic blocks between Tu3 and Ta3B. **c**, A syntenic block composed of five consecutive collinear gene pairs, showing large repeat insertions on Ta3B of >100 kb. **d**, A syntenic block composed of seven consecutive collinear gene pairs. A 70-kb contraction is seen in Ta3B. **e**, A syntenic block with eight collinear genes interrupted by non-syntenic genes. **f**, Two segments of Tu3 and Ta3B show five discrete collinear gene pairs and gene expansions in Tu3. **g**, Genome expansion in one representative syntenic block from Bd2 (0.3 Mb), Tu3 (4.4 Mb), and Ta3B (11.2 Mb). Compared with the Tu3 segment, large numbers of non-homologous genes and repeats can be observed in Ta3B, resulting in a 7-Mb expansion. **h**, Insertion dates of LTR retrotransposons on Tu3 and Ta3B. A recent retrotransposon burst at around 0.1 Ma is observed on Ta3B but not on Tu3.



Extended Data Fig. 9 | Population analysis. a, Distribution of transcriptomic-based SNPs on the seven DNA pseudomolecules (Tu1–Tu7) of *T. urartu*. The SNPs were calculated using a 1-Mb window. **b**, Boxplot comparison of altitude ranges of Tu accessions. The number of accessions in each group is indicated in parentheses. The line inside each box represents the median, the ends of each box define the 25th and 75th percentiles, and the error bars mark the 10th and 90th percentiles. Outliers are displayed as open circles. **c**, Analysis of genetic diversity and differentiation. The π, θ and F_{ST} values were estimated for the three groups of Tu accessions using transcriptomic SNPs. **d**, Reaction phenotypes of Tu accessions to the wheat powdery mildew fungus Bgt. Most of the accessions in Groups I and III (96.7% and 90.6%, respectively) were susceptible to Bgt (race E09), whereas the majority of Group II accessions (92.2%) showed resistance. **e**, A total of 141 (top 1%; $\pi_{\mathrm{Group\ I}}/\pi_{\mathrm{Group\ II}} > 7.7$) and 143 (top 1%; $\pi_{\mathrm{Group\ III}}/\pi_{\mathrm{Group\ II}} > 4.3$) signals were considered to be candidate sweeps (dots above the dashed horizontal threshold line). The Tu accessions in Groups I, II and III were 30, 64 and 53, respectively. Blue arrows indicate the wall-associated receptor protein kinase gene (*TuWAK*, TuG1812G0400002796), whose haplotype variations showed strong associations with resistance or susceptibility to Bgt. **f**, Exon (box)–intron (line) structure of *TuWAK* and its two major haplotypes (Hap1 and Hap2). The positions of the three SNPs that differ between Hap1 and Hap2 are shown. Amino acid changes caused by these SNPs are also displayed. **g**, Distribution of Hap1 and Hap2 in the three groups of *T. urartu* accessions. Hap1 was the major haplotype in Group II, which was strongly associated with resistance to Bgt, while Hap2 was the main haplotype in Groups I and III, and associated with susceptibility to Bgt. In **d** and **g**, the accessions in each group were further sorted based on the response to powdery mildew infection (**d**) or the possession of *TuWAK* haplotypes (**g**), with the assorted accession numbers indicated in appropriate columns.

Extended Data Table 1 | Summary of genome annotations of *T. urartu*

a

Evidence type	Gene#	Percent (%)	Confidence
EST|Protein|cDNA*	23,919	57.63	High
EST|Protein*	10,801	26.02	High
EST|cDNA*	2,196	5.29	High
Protein|cDNA*	600	1.45	High
High-Confidence*	37,516	90.38	High
Protein	2,314	5.57	Low
EST	982	2.37	Low
cDNA	695	1.67	Low
Low-Confidence	3,991	9.62	Low
Total	41,507	100	---

b

	B. distachyon‡	*O. sativa*‡	*T. urartu*
Gene number	31,694	35,472	41,507
Max gene length	47,230	57,648	79,857
Min gene length	90	96	201
Gene length*	2,572/3,298	2,458/3,082	2,329/3,360
mRNA length*	1,428/1,632	1,392/1,557	1,261/1,453
Max CDS length	16,070	16,030	16,086
CDS length*	1,008/1,194	801/991	765/998
Protein length*	336/398	267/330	254/332
Exon length*	138/256	177/358	173/320
Intron length*	139/393	155/438	142/508
5' UTR length*	164/240	133/213	116/212
3' UTR length*	329/396	335/416	257/320
Exon# per transcript†	3.0/4.7	3.0/4.4	3.0/4.5
Transcript# per gene†	1.0/1.4	1.0/1.2	1.0/1.5

c

	Number	Percent (%)
InterPro	30,631	73.80
GO	22,862	55.08
Pathway	4,711	11.35
Pfam	28,837	69.48
Homologous gene#*	31,273	75.34
Annotated	36,602	88.18
Unannotated	4,905	11.82
Total	41,507	100

d

Type	Copy number	Averagel length (bp)	Total length (bp)	Ratio (1e^{-4})
miRNA	31,269	126	3,934,708	8.0946
lncRNA	5,810	320	1,858,475	3.8233
tRNA	3,620	72	261,033	0.537
rRNA				
18S	15	3,483	52,245	0.1075
28S	10	3,420	34,199	0.0704
5.8S	8	158	1,265	0.0026
5S	47	122	5,744	0.0118
snRNA				
CD-box	1,762	143	251,771	0.518
HACA-box	479	130	62,481	0.1285
Splicing	278	147	40,881	0.0841

e

	Percentage of genome (%)					Length (bp)
	Bd	Sb	Os	Zm	Tu	Tu
Class I: Retrotransposon	21.58	50.77	21.00	76.35	71.83	3,442,918,807
LTR-Retrotransposon	18.38	49.70	19.85	75.52	68.61	3,288,676,772
LTR/Gypsy	13.77	42.85	16.39	48.43	42.71	2,047,133,485
LTR/Copia	4.46	6.81	3.08	26.55	24.30	1,164,617,310
Other	0.15	0.04	0.38	0.54	1.61	76,932,977
Non-LTR Retrotransposon	3.20	1.07	1.16	0.84	3.22	154,242,035
SINE	0.26	0.08	0.05	0.03	0.06	2,978,454
LINE	2.94	0.98	1.11	0.80	3.16	151,263,581
Class II: DNA Transposon	5.33	7.17	5.82	5.39	7.41	354,962,893
EnSpm/CACTA	1.44	3.67	2.38	2.06	5.00	239,453,582
hAT	0.43	0.26	0.27	0.75	0.77	36,758,465
Harbinger	0.26	0.20	0.08	0.22	0.56	26,610,774
Tc1/Mariner	1.19	0.61	0.03	0.07	0.02	1,048,894
Helitron	0.06	0.13	0.00	0.54	0.27	12,725,584
Other	3.48	3.05	4.44	1.43	0.80	38,372,594
Tandem repeat	1.89	2.49	2.90	0.86	1.21	58,076,221
Low complexity	0.27	0.19	0.82	0.12	0.07	3,195,334
Unclassified	8.41	5.21	0.23	0.74	0.91	43,547,798
Total content	37.48	65.83	30.78	82.48	81.42	3,902,708,053

a, The number of genes supported by each evidence type in core gene set of *T. urartu*. *Vertical bar indicates multiple types of evidence support. Cross-species evidence was categorized as protein evidence. **b**, Gene numbers and features of *T. urartu*. *The median/average length in base pair (bp). †The median/average in number. ‡*B. distachyon*: Bdistachyon_283_v2.1 gene set. *O. sativa*: RAPV1 gene set. **c**, Functional annotation of the predicted genes in *T. urartu*. *Based on comparison with *B. distachyon* and *O. sativa*. **d**. Summary of predicted non-coding RNAs in the genome of *T. urartu*. **e**, Comparison of repetitive DNA content among *T. urartu* and four other sequenced grasses. Bd, *B. distachyon*; Sb, *S. bicolor*; Os, *O. sativa*; Zm, *Z. mays*, Tu, *T. urartu*.

LETTER

https://doi.org/10.1038/s41586-018-0102-6

Vector-based navigation using grid-like representations in artificial agents

Andrea Banino[1,2,3,5]*, Caswell Barry[2,5]*, Benigno Uria[1], Charles Blundell[1], Timothy Lillicrap[1], Piotr Mirowski[1], Alexander Pritzel[1], Martin J. Chadwick[1], Thomas Degris[1], Joseph Modayil[1], Greg Wayne[1], Hubert Soyer[1], Fabio Viola[1], Brian Zhang[1], Ross Goroshin[1], Neil Rabinowitz[1], Razvan Pascanu[1], Charlie Beattie[1], Stig Petersen[1], Amir Sadik[1], Stephen Gaffney[1], Helen King[1], Koray Kavukcuoglu[1], Demis Hassabis[1,4], Raia Hadsell[1] & Dharshan Kumaran[1,3]*

Deep neural networks have achieved impressive successes in fields ranging from object recognition to complex games such as Go[1,2]. Navigation, however, remains a substantial challenge for artificial agents, with deep neural networks trained by reinforcement learning[3–5] failing to rival the proficiency of mammalian spatial behaviour, which is underpinned by grid cells in the entorhinal cortex[6]. Grid cells are thought to provide a multi-scale periodic representation that functions as a metric for coding space[7,8] and is critical for integrating self-motion (path integration)[6,7,9] and planning direct trajectories to goals (vector-based navigation)[7,10,11]. Here we set out to leverage the computational functions of grid cells to develop a deep reinforcement learning agent with mammal-like navigational abilities. We first trained a recurrent network to perform path integration, leading to the emergence of representations resembling grid cells, as well as other entorhinal cell types[12]. We then showed that this representation provided an effective basis for an agent to locate goals in challenging, unfamiliar, and changeable environments—optimizing the primary objective of navigation through deep reinforcement learning. The performance of agents endowed with grid-like representations surpassed that of an expert human and comparison agents, with the metric quantities necessary for vector-based navigation derived from grid-like units within the network. Furthermore, grid-like representations enabled agents to conduct shortcut behaviours reminiscent of those performed by mammals. Our findings show that emergent grid-like representations furnish agents with a Euclidean spatial metric and associated vector operations, providing a foundation for proficient navigation. As such, our results support neuroscientific theories that see grid cells as critical for vector-based navigation[7,10,11], demonstrating that the latter can be combined with path-based strategies to support navigation in challenging environments.

The ability to self-localize in the environment and to update one's position on the basis of self-motion are core components of navigation[13]. We trained a deep neural network to path integrate within a square arena (2.2 m × 2.2 m), using simulated trajectories modelled on those of foraging rodents (see Methods). The network was required to update its estimate of location and head direction using translational and angular velocity signals, mirroring those available to the mammalian brain[12,14,15] (see Methods; Fig. 1a, b). Velocity was provided as input to a recurrent network with a long short-term memory (LSTM) architecture, which was trained using backpropagation through time (see Methods and Supplementary Discussion), allowing the network to dynamically combine current input signals with activity patterns reflecting past events (see Methods; Fig. 1a). The LSTM projected to place and head direction units via a linear layer—units with activity defined as a simple linear function of their input (see Extended Data Fig. 1 for architecture). Importantly, the linear layer was subject to regularization, in particular dropout[16], such that 50% of the units were randomly silenced at each time step. The vector of activities in the place and head direction units, corresponding to the current position, was provided as a supervised training signal at each time step (see Methods; Extended Data Fig. 1). This form of supervision follows evidence that in mammals, place and head direction representations exist in close anatomical proximity to entorhinal grid cells[12] and emerge in rodent pups before the appearance of mature grid cells[17,18]. Equally, in adult rodents, entorhinal grid cells are known to project to the hippocampus and appear to contribute to the neural activity of place cells[19].

As expected, the network was able to path integrate accurately in this setting involving foraging behaviour (mean error after 15 s trajectory was 16 cm versus 91 cm for an untrained network, effect size 2.83, 95% confidence interval (CI) 2.80–2.86; Fig. 1b, c). Strikingly, individual units within the linear layer of the network developed stable spatial activity profiles similar to those of neurons within the entorhinal network[6,12] (Fig. 1d, Extended Data Fig. 2). Specifically, 129 of the 512 linear layer units (25.2%) resembled grid cells, exhibiting significant hexagonal periodicity (gridness[20]) versus a null distribution generated by a conservative field shuffling procedure (see Methods). The scale of the grid-patterns, measured from the spatial autocorrelograms of the activity maps[20], varied between units (range 28 cm to 115 cm, mean 66 cm) and followed a multi-modal distribution, consistent with empirical results from rodent grid cells[21,22] (Fig. 1e). To assess these clusters, we fit mixtures of Gaussians, finding the most parsimonious number by minimizing the Bayesian information criterion (BIC). The distribution was best fit by three Gaussians (means 47 cm, 70 cm, and 106 cm), indicating the presence of scale clusters with a ratio between neighbouring clusters of approximately 1.5, closely matching theoretical predictions[23] and also lying within the range reported for rodents[21,22] (Fig. 1e, Extended Data Fig. 3). The linear layer also exhibited units resembling head direction cells (10.2%), border cells (8.7%), and a small number of place cells[12] as well as conjunctions of these representations[20,24] (Fig. 1d, f, g, Extended Data Fig. 2).

To ascertain how robust these representations were, we retrained the network 100 times, in each instance finding similar proportions of grid-like units (mean 23%, s.d. 2.8%, units with significant gridness scores) and other spatially modulated units (Extended Data Fig. 3). Conversely, grid-like representations did not emerge in networks without regularization (for example, dropout, see Methods; also see ref. 25, Extended Data Fig. 4). Therefore, the use of regularization, including dropout (which has been viewed to be a parallel of noise in neural systems[16]), was critical to the emergence of entorhinal-like representations. Notably, therefore, our results show that grid-like representations reminiscent of those found in the mammalian entorhinal cortex emerge in a generic network trained to path integrate, contrasting with previous approaches using pre-configured grid cells[26] (see Supplementary Discussion).

---

[1]DeepMind, London, UK. [2]Department of Cell and Developmental Biology, University College London, London, UK. [3]Centre for Computation, Mathematics and Physics in the Life Sciences and Experimental Biology (CoMPLEX), University College London, London, UK. [4]Gatsby Computational Neuroscience Unit, University College London, London, UK. [5]These authors contributed equally: Andrea Banino, Caswell Barry. *e-mail: abanino@google.com; caswell.barry@ucl.ac.uk; dkumaran@google.com

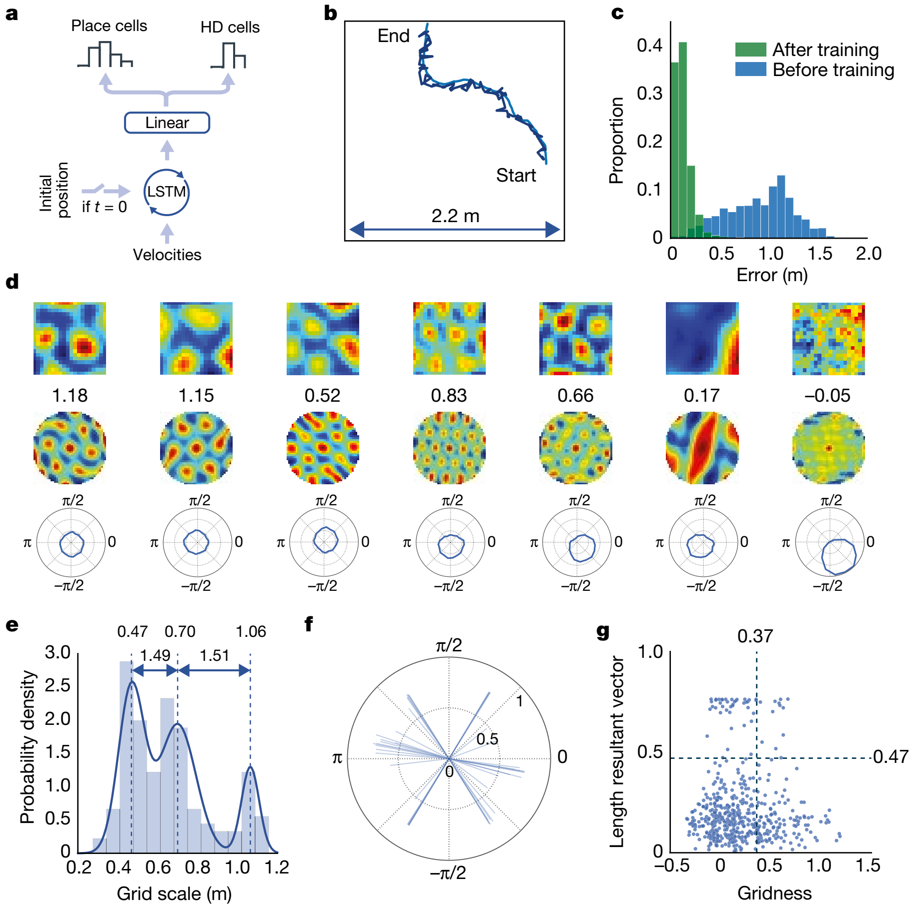


Fig. 1 | Entorhinal-like representations emerge in a network trained to path integrate. **a**, Schematic of network architecture (see Extended Data Fig. 1). **b**, Example trajectory (15 s); self-location decoded from place cells (dark blue) resembles actual path (light blue). **c**, Accuracy of decoded location before (blue) and after (green) training. **d**, Linear layer units exhibit spatially tuned responses resembling grid, border, and head direction cells. Top, ratemap shows activity over location; middle, spatial autocorrelogram of ratemap with gridness indicated; bottom, polar plot show activity versus head direction. **e**, Spatial scale of grid-like units ($n = 129$) is clustered. Distribution is more discrete than by chance (effect size, 2.98; 95% CI, 0.97–4.91) and best fit by a mixture of three Gaussians (centres 0.47, 0.70 and 1.06 m, ratios are 1.49 and 1.51). **f**, Directional tuning of the most strongly directional units ($n = 52$). Lines indicate length and orientation of resultant vector (see Methods), exhibiting six-fold clustering reminiscent of conjunctive grid cells. **g**, Distribution of gridness and directional tuning. Dashed lines indicate 95% confidence interval from null distributions (based on 500 data permutations); 14 (11%) grids exhibit directional modulation (see Methods). Similar results were seen in a circular environment (Extended Data Fig. 3).

Furthermore, our results are consistent with the view that grid cells represent an efficient and robust basis for a location code updated by self-motion cues[6–9].

Next, we sought to test the hypothesis that the emergent representations provide an effective basis function for goal-directed navigation in novel challenging and changeable environments, when trained through deep reinforcement learning. Entorhinal grid cells have been proposed to provide a Euclidean spatial metric and thus to support the calculation of goal-directed vectors, enabling animals to follow direct routes to a remembered goal, a process known as vector-based navigation[7,10,11]. Theoretically, the advantage of decomposing spatial location into a multi-scale periodic code, as provided by grid cells, is that the relative positions of two points can be retrieved by examining the difference in the code at the level of each scale—combining the modulus remainders to return the true vector[7,11] (Fig. 2a). However, despite the obvious utility of such a framework, experimental evidence for the direct involvement of grid representations in goal-directed navigation is still lacking.

To develop an agent with the potential for vector-based navigation, we incorporated the 'grid network' described above into a larger architecture that was trained using deep reinforcement learning (Fig. 2d, Extended Data Fig. 5). As before, the grid network was trained using supervised learning but, to better approximate the information available to navigating mammals, it now received velocity signals perturbed with random noise as well as visual input. Experimental evidence suggests that place cell input to grid cells corrects for drift and anchors grids to environmental cues[21]. To parallel this, visual input was processed by a 'vision module' consisting of a convolutional network that produced place and head direction cell activity patterns that were provided as input to the grid network 5% of the time—akin to a moving animal making occasional, imperfect observations of salient environmental cues[27] (see Methods; Fig. 2b, c and Extended Data Fig. 5). The output of the linear layer of the grid network, corresponding to the agent's current location, was provided as input to the 'policy LSTM', a second recurrent network that both controls the agent's actions and outputs a value function. Additionally, whenever the agent reached the goal, the 'goal grid code'—activity in the linear layer—was subsequently provided to the policy LSTM during navigation as an additional input.

We first examined the navigational capacities of the agent in a simple setting inspired by the classic Morris water maze (Fig. 2b, c; 2.5 m × 2.5 m square arena; see Methods and Supplementary Results). Notably, the agent was still able to self-localize accurately in this more challenging setting, where ground truth information about location was not provided and velocity inputs were noisy (mean error after 15-s trajectory, 12 cm versus 88 cm for an untrained network; effect size, 2.82; 95% CI, 2.79–2.84; Fig. 2e). Furthermore, the agent exhibited proficient goal-finding abilities, typically taking direct routes to the goal from arbitrary starting locations (Fig. 2h). Performance exceeded that of a control place cell agent (Fig. 2f, see Supplementary Results and

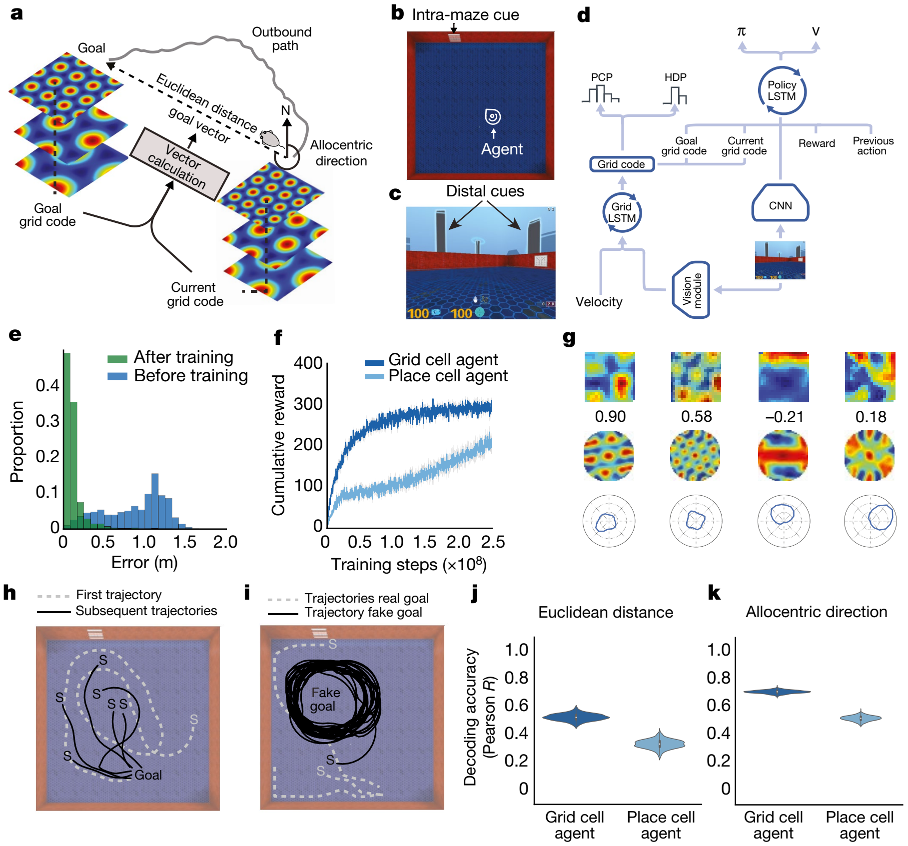


Fig. 2 | One-shot open field navigation to a hidden goal. a, Schematic of vector-based navigation. **b**, Overhead view of typical environment (icon indicates agent and facing direction). **c**, Agent view of **b**. **d**, Schematic of deep reinforcement learning architecture (Extended Data Fig. 5). PCP, place cell predictions; HDP, head direction cell predictions. **e**, Accuracy of self-location decoded from place cell units. **f**, Performance of grid cell agent and place cell agent (y axis shows reward obtained within a single episode, 10 points per goal arrival, grey band displays the 68% confidence interval based on 5,000 bootstrapped samples). **g**, As in Fig. 1, the linear layer develops spatial representations similar to entorhinal cortex. Left to right, two grid cells, one border cell, and one head direction cell. **h**, On the first trial of an episode, the agent explores to find the goal and subsequently navigates directly to it. 'S' denotes the starting location. **i**, After successful navigation, the policy LSTM was supplied with a 'fake' goal grid-code, directing the agent to this location where no goal was present. **j**, **k**, Decoding of goal-directed metric codes (that is, Euclidean distance and direction) from the policy LSTM of grid cell and place cell agents. The bootstrapped distribution (1,000 samples) of correlation coefficients are each displayed with a violin plot overlaid on a Tukey boxplot.

Methods), chosen because place cells provide a robust representation of self-location but are not thought to provide a substrate for long-range vector calculations[11]. We examined the units in the linear layer, again finding a heterogeneous population resembling those found in entorhinal cortex, including grid-like units (21.4%) as well as other spatial representations (Fig. 2g, Extended Data Fig. 6)—paralleling the dependence of mammalian grid cells on self-motion information[15,28] and spatial cues[6,21].

We next turn to our central claim, that grid cells endow agents with the ability to perform vector-based navigation, enabling downstream regions to calculate goal-directed vectors by comparing current activity with that of a remembered goal[7,10,11]. In the agent, we expect these calculations to be performed by the policy LSTM, which receives the current activity pattern over the linear layer (termed 'current grid code; Fig. 2d and Extended Data Fig. 5) as well as that present the last time the agent reached the goal (termed 'goal grid code') and uses them to control movement. Hence we performed several manipulations, which yielded four lines of evidence in support of the vector-based navigation hypothesis (see Supplementary Results).

First, to demonstrate that the goal grid code provided sufficient information to enable the agent to navigate to an arbitrary location, we substituted it with a 'fake' goal grid code sampled randomly from a location in the environment (see Methods). The agent followed a direct path to the newly specified location, circling the absent goal (Fig. 2i)—similar to rodents in probe trials of the Morris water maze (escape platform removed). Second, we demonstrated that withholding the goal grid code from the policy LSTM of the grid cell agent had a strikingly deleterious effect on performance (Extended Data Fig. 6c). Third, we demonstrated that the policy LSTM of the grid cell agent contained representations of key components of vector-based navigation (Fig. 2j, k), and that both Euclidean distance (difference in $r = 0.17$; 95% CI, 0.11–0.24) and allocentric goal direction (difference in $r = 0.22$; 95% CI, 0.18–0.26) were represented more strongly than in the place cell agent. Notably, a neural representation of goal distance has recently been reported in the mammalian hippocampus[29]. Finally, we provide evidence consistent with a prediction of the vector-based navigation hypothesis, namely that a targeted lesion (that is, silencing) to the most grid-like units within the goal grid code should have a greater adverse effect on performance and the representation of vector-based metrics (for example, Euclidean distance) than a sham lesion (that is, silencing of non-grid units; see Supplementary Results).

Having demonstrated the effectiveness of grid-like representations in optimizing one-shot goal learning in a simple square arena, we assessed the agent's performance in two challenging, procedurally generated multi-room environments, referred to as 'goal-driven' and 'goal-doors' (see Methods). Notably, these environments are challenging for deep reinforcement learning agents with external memory (Extended Data Fig. 7e, f, h, i and Supplementary Results). Again, the grid cell agent exhibited high levels of performance, was strikingly robust across a range of network hyperparameters (Extended Data Fig. 7a–c), and reached the goal more frequently than either control agents or a human expert—a typical benchmark for the performance of deep

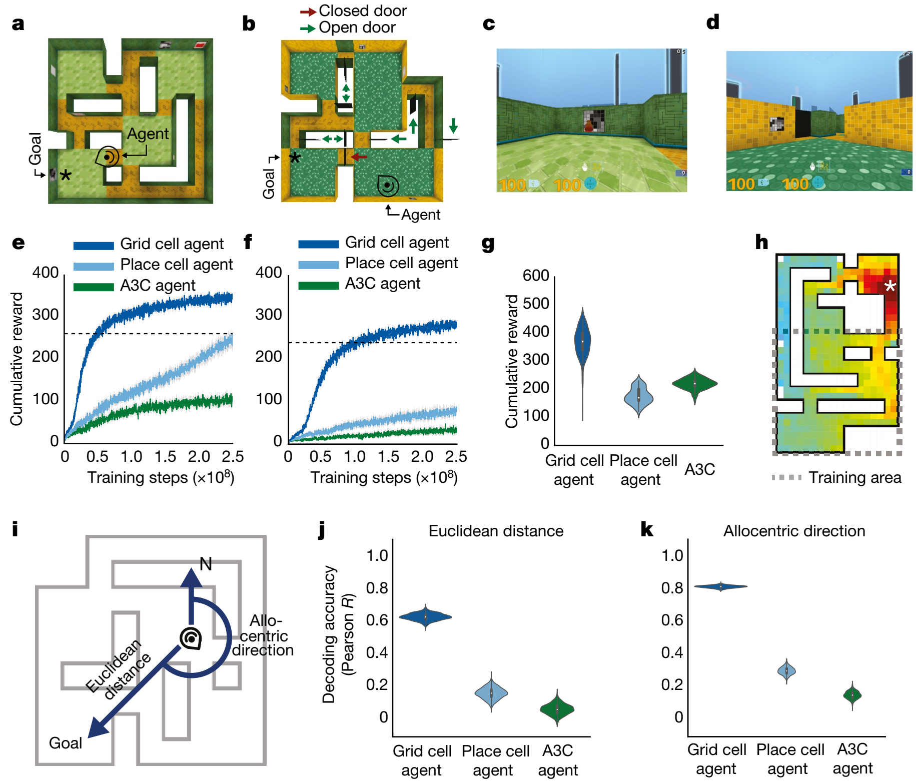


Fig. 3 | Navigation in challenging environments. **a**, **b**, Overhead view of multi-room environments goal-driven (**a**) and goal-doors (that is, with stochastic doors; **b**). Goal (asterisk) and agent (head icon) locations are displayed. **c**, **d**, Agent views of **a** and **b**, respectively, showing red goal and closed black door. **e**, **f**, Agent training performance curves for **a** and **b**, respectively, and performance of human expert (dashed line). Performance is mean cumulative reward over 100 episodes. The grey band displays the 68% CI based on 5,000 bootstrapped samples. **g**, Distribution of test performance over 100 episodes, showing ability of agents to generalize to a larger version of the goal-doors environment, displayed with a violin plot overlaid on a Tukey boxplot for each agent. **h**, The value function of the grid cell agent is projected onto an example larger goal doors environment as a heat map (where red indicates higher value function). Dotted lines show the extent of the original training environment. Despite the larger size, the value function clearly approximates Euclidean distance to goal. **i**, Schematic displaying the key metrics required for vector-based navigation to a goal. **j**, **k**, Decoding of vector-based metric codes from the policy LSTMs of agents during navigation. The bootstrapped distribution (1,000 samples) of correlation coefficients are displayed with a violin plot overlaid on a Tukey boxplot in each case.

reinforcement learning agents in game playing scenarios[2] (Fig. 3e, f; see Supplementary Results). Furthermore, when agents were tested without retraining in environments considerably larger than those seen previously, only the grid cell agent was able to generalize effectively (Fig. 3g, h; see Supplementary Results). Despite the complexity of the 'goal-driven' environment, we could still decode the key metric codes from the grid agent policy LSTM with high accuracy during the initial period of navigation, and decoding accuracy was substantially higher in the grid cell agent than in both the place cell and deep reinforcement learning control agents (Fig. 3j, k and Supplementary Results; see Extended Data Figs. 8, 9 for control agent architectures).

Finally, a core feature of mammalian spatial behaviour is the ability to exploit novel shortcuts and traverse unvisited portions of space, a capacity thought to depend on vector-based navigation[9,11]. Strikingly, the grid cell agent—but not comparison agents—robustly demonstrated these abilities in specifically designed neuroscience-inspired mazes, taking direct routes to the goal as soon as they became available (Fig. 4, Extended Data Fig. 10 and Supplementary Results).

Conventional simultaneous localization and mapping (SLAM) techniques typically require an accurate and complete map to be built, with the nature and position of the goal externally defined[30]. By contrast, the deep reinforcement learning approach described in this work has the ability to learn complex control policies end-to-end from a sparse reward, taking direct routes involving shortcuts to goals in an automatic fashion—abilities that exceed previous deep reinforcement learning approaches[3–5], and that would have to be hand-coded in any SLAM system.

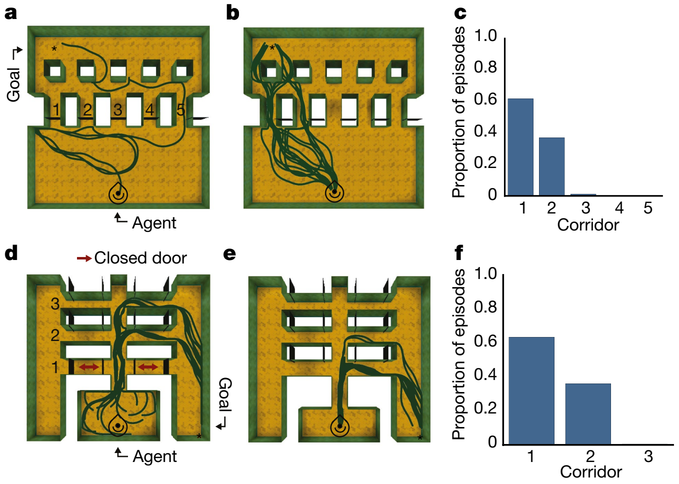


Fig. 4 | Flexible use of shortcuts. **a**, Example trajectory from grid cell agent in the linear sunburst maze (only door 5 open; icon indicates start location). **b**, Testing configuration with all doors open; grid cell agent uses the newly available shortcuts (100 episodes shown). **c**, Histogram showing agent's strong preference for most direct routes. **d**, Example grid cell agent trajectories (100) during training in the double E-maze (corridor 1 doors closed). **e**, Testing configuration with corridor 1 open, and 100 grid agent trajectories. **f**, Histogram analogous to **c** showing that agent prefers newly available shortest route. See Extended Data Fig. 10 for performance of place cell agent.

Our work, in demonstrating that grid-like representations provide an effective basis for flexible navigation in challenging novel environments, supports theoretical models of grid cells in vector-based navigation that were previously lacking strong empirical support[7,10,11]. We also show that vector-based navigation can be effectively combined with a path-based barrier avoidance strategy to enable the exploitation of optimal routes in challenging multi-compartment environments. In sum, we argue that grid-like representations furnish agents with a Euclidean geometric framework—paralleling their proposed computational role in mammals as an early-developing Kantian-like spatial scaffold that serves to organize perceptual experience[17,18].

Online content

Any Methods, including any statements of data availability and Nature Research reporting summaries, along with any additional references and Source Data files, are available in the online version of the paper at https://doi.org/10.1038/s41586-018-0102-6.

Received: 5 July 2017; Accepted: 3 April 2018;
Published online 9 May 2018.

1. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436–444 (2015).
2. Silver, D. et al. Mastering the game of Go with deep neural networks and tree search. *Nature* **529**, 484–489 (2016).
3. Oh, J., Chockalingam, V., Singh, S. P. & Lee, H. Control of memory, active perception, and action in Minecraft. *Proc. Intl Conf. Machine Learning* **48** (2016).
4. Kulkarni, T. D., Saeedi, A., Gautam, S. & Gershman, S. J. Deep successor reinforcement learning. Preprint at https://arxiv.org/abs/1606.02396 (2016).
5. Mirowski, P. et al. Learning to navigate in complex environments. *Intl Conf. Learning Representations* (2017).
6. Hafting, T., Fyhn, M., Molden, S., Moser, M.-B. & Moser, E. I. Microstructure of a spatial map in the entorhinal cortex. *Nature* **436**, 801–806 (2005).
7. Fiete, I. R., Burak, Y. & Brookings, T. What grid cells convey about rat location. *J. Neurosci.* **28**, 6858–6871 (2008).
8. Mathis, A., Herz, A. V. & Stemmler, M. Optimal population codes for space: grid cells outperform place cells. *Neural Comput.* **24**, 2280–2317 (2012).
9. McNaughton, B. L., Battaglia, F. P., Jensen, O., Moser, E. I. & Moser, M.-B. Path integration and the neural basis of the 'cognitive map'. *Nat. Rev. Neurosci.* **7**, 663–678 (2006).
10. Erdem, U. M. & Hasselmo, M. A goal-directed spatial navigation model using forward trajectory planning based on grid cells. *Eur. J. Neurosci.* **35**, 916–931 (2012).
11. Bush, D., Barry, C., Manson, D. & Burgess, N. Using grid cells for navigation. *Neuron* **87**, 507–520 (2015).
12. Barry, C. & Burgess, N. Neural mechanisms of self-location. *Curr. Biol.* **24**, R330–R339 (2014).
13. Mittelstaedt, M.-L. & Mittelstaedt, H. Homing by path integration in a mammal. *Naturwissenschaften* **67**, 566–567 (1980).
14. Bassett, J. P. & Taube, J. S. Neural correlates for angular head velocity in the rat dorsal tegmental nucleus. *J. Neurosci.* **21**, 5740–5751 (2001).
15. Kropff, E., Carmichael, J. E., Moser, M.-B. & Moser, E. I. Speed cells in the medial entorhinal cortex. *Nature* **523**, 419–424 (2015).
16. Srivastava, N., Hinton, G. E., Krizhevsky, A., Sutskever, I. & Salakhutdinov, R. Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.* **15**, 1929–1958 (2014).
17. Wills, T. J., Cacucci, F., Burgess, N. & O'Keefe, J. Development of the hippocampal cognitive map in preweanling rats. *Science* **328**, 1573–1576 (2010).
18. Langston, R. F. et al. Development of the spatial representation system in the rat. *Science* **328**, 1576–1580 (2010).
19. Zhang, S.-J. et al. Optogenetic dissection of entorhinal-hippocampal functional connectivity. *Science* **340**, 1232627 (2013).
20. Sargolini, F. et al. Conjunctive representation of position, direction, and velocity in entorhinal cortex. *Science* **312**, 758–762 (2006).
21. Barry, C., Hayman, R., Burgess, N. & Jeffery, K. J. Experience-dependent rescaling of entorhinal grids. *Nat. Neurosci.* **10**, 682–684 (2007).
22. Stensola, H. et al. The entorhinal grid map is discretized. *Nature* **492**, 72–78 (2012).
23. Stemmler, M., Mathis, A. & Herz, A. V. Connecting multiple spatial scales to decode the population activity of grid cells. *Sci. Adv.* **1**, e1500816 (2015).
24. Doeller, C. F., Barry, C. & Burgess, N. Evidence for grid cells in a human memory network. *Nature* **463**, 657–661 (2010).
25. Kanitscheider, I. & Fiete, I. Training recurrent networks to generate hypotheses about how the brain solves hard navigation problems. Preprint at https://arxiv.org/abs/1609.09059 (2016).
26. Milford, M. J. & Wyeth, G. F. Mapping a suburb with a single camera using a biologically inspired slam system. *IEEE Trans. Robot.* **24**, 1038–1053 (2008).
27. Hardcastle, K., Ganguli, S. & Giocomo, L. M. Environmental boundaries as an error correction mechanism for grid cells. *Neuron* **86**, 827–839 (2015).
28. Chen, G., King, J. A., Burgess, N. & O'Keefe, J. How vision and movement combine in the hippocampal place code. *Proc. Natl Acad. Sci. USA* **110**, 378–383 (2013).
29. Sarel, A., Finkelstein, A., Las, L. & Ulanovsky, N. Vectorial representation of spatial goals in the hippocampus of bats. *Science* **355**, 176–180 (2017).
30. Dissanayake, M. G., Newman, P., Clark, S., Durrant-Whyte, H. F. & Csorba, M. A solution to the simultaneous localization and map building (slam) problem. *IEEE Trans. Robot. Autom.* **17**, 229–241 (2001).

Acknowledgements We thank M. Jaderberg, V. Mnih, A. Santoro, T. Schaul, K. Stachenfeld and J. Yosinski for discussions, and M. Botvinick and J. Wang for comments on an earlier version of the manuscript. C.Ba. funded by Royal Society and Wellcome Trust.

Reviewer information *Nature* thanks J. Conradt and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions Conceived project: A.B., D.K., C.Ba., R.H., P.M. and B.U.; contributed ideas to experiments: A.B., D.K., C.Ba., B.U., R.H., T.L., C.Bl., P.M., A.P., T.D., J.M., K.K., N.R., G.W., R.G., M.J.C., D.H. and R.P.; performed experiments and analysis: A.B., C.Ba., B.U., M.J.C., T.L., H.S., A.P., B.Z. and F.V.; development of testing platform and environments: C.Be., S.P., R.H., T.L., G.W., D.K., A.B., B.U. and D.H.; human expert tester: A.S.; managed project: D.K., R.H., A.B., H.K., S.G. and D.H.; wrote paper; D.K., A.B., C.Ba., T.L., C.Bl., B.U., M.C., A.P., R.H., N.R., K.K. and D.H.

Competing interests The authors declare no competing interests.

Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0102-6.
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0102-6.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to A.B. or C.B. or D.K.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

Path integration: supervised learning experiments. *Simplified 2D environment.* Simulated rat trajectories of duration T were generated in square and circular environments with walls of length L (diameter in the circular case). The simulated rat started at a uniformly sampled location and facing angle within the enclosure. A rat-like motion model[31] was used to obtain trajectories that uniformly covered the whole environment by avoiding walls (see Supplementary Methods Table 1 for the model's parameters).

Ground truth place cell distribution. Place cell activations, $\overrightarrow{c} \in [0, 1]^N$, for a given position $\overrightarrow{x} \in R^2$ were simulated by the posterior probability of each component of a mixture of two- dimensional isotropic Gaussians,

$$c_i = \frac{e^{-\frac{\|\overrightarrow{x} - \overrightarrow{\mu}_i^{(c)}\|_2^2}{2(\sigma^{(c)})^2}}}{\sum_{j=1}^{N} e^{-\frac{\|\overrightarrow{x} - \overrightarrow{\mu}_j^{(c)}\|_2^2}{2(\sigma^{(c)})^2}}}$$

where $\overrightarrow{\mu}_i^{(c)} \in \mathbb{R}^2$, the place cell centres, are N two-dimensional vectors chosen uniformly at random before training, and $\sigma^{(c)}$, the place cell scale, is a positive scalar fixed for each experiment.

Ground truth head-direction cell distribution. Head-direction cell activations, $\overrightarrow{h} \in [0, 1]^M$, for a given facing angle φ were represented by the posterior probability of each component of a mixture of Von Mises distributions with concentration parameter $\kappa^{(h)}$,

$$h_i = \frac{e^{k^{(h)} \cos(\varphi - \mu_i^{(h)})}}{\sum_{j=1}^{M} e^{k^{(h)} \cos(\varphi - \mu_j^{(h)})}}$$

where the M head direction centres $\mu_i^{(h)} \in [-\pi, \pi]$ are chosen uniformly at random before training, and $\kappa^{(h)}$, the concentration parameter, is a positive scalar fixed for each experiment.

Supervised learning inputs. In the supervised setup the grid cell network receives, at each step t, the egocentric linear velocity $v_t \in \mathbb{R}$ and the sine and cosine of its angular velocity φ_t.

Grid cell network architecture. The grid cell network architecture (Extended Data Fig. 1) consists of three layers: a recurrent layer, a linear layer, and an output layer. The single recurrent layer is an LSTM (long short-term memory[32]) that projects to place and head direction units via the linear layer. The linear layer implements regularization through dropout[16]. The recurrent LSTM layer consists of one cell of 128 hidden units, with no peephole connections. Input to the recurrent LSTM layer is the vector $[v_t, sin(\varphi_t), cos(\varphi_t)]$. The initial cell state and hidden state of the LSTM, $\overrightarrow{l}_0$ and $\overrightarrow{m}_0$, respectively, are initialized by computing a linear transformation of the ground truth place and head-direction cells at time 0,

$$\overrightarrow{l}_0 = W^{cp} \overrightarrow{c}_0 + W^{(cd)} \overrightarrow{h}_0$$
$$\overrightarrow{m}_0 = W^{hp} \overrightarrow{c}_0 + W^{(hd)} \overrightarrow{h}_0$$

The parameters of these two linear transformations ($W^{(cp)}$, $W^{(cd)}$, $W^{(hp)}$ and $W^{(hd)}$) were optimized during training. The output of the LSTM, $\overrightarrow{m}_t$ is then used to produce predictions of the place cells $\overrightarrow{y}_t$ and head direction cells $\overrightarrow{z}_t$ by means of a linear decoder network.

The linear decoder consists of three sets of weights and biases. The first set consists of the weights and biases that map from the LSTM hidden state $\overrightarrow{m}_t$ to the linear layer activations $\overrightarrow{g}_t \in \mathbb{R}^{512}$. The other two sets of weights map from the linear layer activations $\overrightarrow{g}_t$ to the predicted head directions, $\overrightarrow{z}_t$ and predicted place cells, $\overrightarrow{y}_t$, respectively, via softmax functions[33]. Dropout[16] with drop probability 0.5 was applied to each $\overrightarrow{g}_t$ unit. Note that there is no intermediary nonlinearity in the linear decoder.

Supervised learning loss. The grid cell network is trained to predict the place and head-direction cell ensemble activations, $\overrightarrow{c}_t$ and $\overrightarrow{h}_t$, respectively, at each time step t. During training, the network was trained in a single environment where the place cell centres were constant throughout. The parameters of the grid cell network are trained by minimizing the cross-entropy between the network place cell predictions, $\overrightarrow{y}_t$, and the synthetic place-cells targets, $\overrightarrow{c}_t$, and the cross-entropy between head-direction predictions, $\overrightarrow{z}_t$, and their targets, $\overrightarrow{h}_t$,

$$\mathcal{L}(\overrightarrow{y}, \overrightarrow{z}, \overrightarrow{c}, \overrightarrow{h}) = -\sum_{i=1}^{N} c_i \log(y_i) - \sum_{j=1}^{N} h_j \log(z_j)$$

Gradients of this loss function with respect to the network parameters were calculated using backpropagation through time[34], unrolling the network into blocks of 100 time steps. The network parameters were updated using stochastic-gradient descent (RMSProp[35]), with weight decay[36] for the weights projecting from the dropout linear layer, $\overrightarrow{g}_t$, to the place and head-direction cell predictions, $\overrightarrow{y}_t$ and $\overrightarrow{z}_t$. Hyperparameter values used for training are listed in Supplementary Methods Table 1.

Gradient clipping. In our simulations, gradient clipping was used for parameters projecting from the dropout linear layer, $\overrightarrow{g}_t$, to the place and head-direction cell predictions $\overrightarrow{y}_t$ and $\overrightarrow{z}_t$. Gradient clipping clips each element of the gradient vector to lie in a given interval $[-g_c, g_c]$, and is an important tool for optimization in deep and recurrent artificial neural networks, where it helps by preventing exploding gradients[37]. Gradient clipping also introduces distortions into the weight updates, which help to avoid local minima[38].

Navigation through deep reinforcement learning. *Environments and task.* We assessed the performance of agents on three environments seen by the agent from a first-person perspective in the DeepMind Laboratory[39] platform.

Custom environment: square arena. This comprised a 10 × 10 square arena, which corresponds to 2.5 × 2.5 m arena assuming an agent speed of 15 cm/s (Fig. 2b, c). The arena contained a single, coloured, intra-arena cue whose position and colour changed on each episode, as did the texture of the floor, the texture of the walls and the goal location. As in the goal-driven and goal-door environments described below, there were a set of distal cues (that is, buildings) that paralleled the design of virtual reality environments used in human experiments[40]. These distal cues were rendered at infinity—so as to provide directional but not distance information—and their configuration was consistent across episodes. At the start of each episode the agent (described below) started in a random location and was required to explore in order to find an unmarked goal, paralleling the task of rodents in the classic Morris water maze. The agent always started in the central 6 × 6 grids (that is, 1.5 × 1.5 m) of the environment. Noise in the velocity input $\overrightarrow{u}_t$ was applied throughout training and testing (that is, Gaussian noise ε, with $\mu = 0$ and $\sigma = 0.01$). The action space is discrete (six actions) but affords fine-grained motor control (that is, the agent could rotate in small increments, accelerate forwards, backwards or sideways, or effect rotational acceleration while moving).

DeepMind laboratory environments: goal-driven and goal-doors. Goal-driven and goal-doors are challenging, visually rich multi-room environments (Fig. 3a–d). Mazes were formed within an 11 × 11 grid, corresponding to 2.7 × 2.7 m (see below for definition of larger 11 × 17 mazes). Mazes were procedurally generated at the beginning of each episode; thus, the layout, wall texture, landmarks (intra-maze cues on walls) and goal location were different for each episode but consistent within an episode. Distal cues, in the form of buildings rendered at infinity, were as described above for the square arena.

The critical difference between goal-driven and goal-doors tasks is that the latter had the additional challenge of stochastic doors within the maze. Specifically, the state of the doors (open or closed) changed randomly during an episode each time the agent reached the goal. This meant that the optimal path to the goal from a given location changed during an episode, requiring the agent to recompute trajectories.

In both tasks, the agent starts at a random location within the maze and its task is to explore to find the goal. The goal in both levels was always represented by the same object (Fig. 3c). After getting to the goal the agent received a reward of 10 points, after which it was teleported to a new random location within the maze. In both levels, episodes lasted a fixed duration of 5,400 environment steps (90 s).

Generalization on larger environments. We tested the ability of agents trained on the standard environment (11 × 11) to generalize to larger environments (11 × 17, corresponding to 2.7 × 4.25 m). The procedural generation and composition of these environments was done as for the standard environments. Each agent was trained in the 11 × 11 goal-doors maze for a total of 10^9 environment steps, and the best performing replica (highest asymptotic performance averaged over 100 episodes in 11 × 11) was selected for evaluation in the larger maze. Note that the weights of the agent were frozen during evaluation on the larger maze. Evaluation was over 100 episodes of fixed duration 12,600 environment steps (210 s).

Probe mazes to assess shortcut behaviour. To test the agent's ability to follow novel, goal- directed routes, we created a series of environments inspired by mazes designed to test the shortcut abilities of rodents.

The first maze is a linearized version of Tolman's sunburst maze (Fig. 4a) used to determine whether the agent was able to follow an accurate heading towards the goal when a path became available (see Supplementary Methods). In this maze, after reaching the goal, the agent was teleported to the original position with the same heading orientation. Agents were trained in the goal-doors maze and network weights were frozen during testing—all the agents were tested for 100 episodes, each one lasting for a fixed duration of 5,400 environment steps (90 s).

The second environment, the double E-maze (Fig. 4d, Extended Data Fig. 10), was designed to test the agent's ability to traverse an entirely new portion of space (see Supplementary Methods). In this maze we had both training and testing conditions. Both training and testing were conducted within the maze but at test time weights were frozen. The agent always started in the central room (for example, see Fig. 4d). The maze had stochastic doors with two different configurations, one for the training phase and one for testing phase. During training, the state of the doors (open or closed) randomly changed during an episode each time the agent reached the goal. Critically, during training the corridors presenting the shortest route to the goal (that is, the ones closer to the central room) were closed at both ends, preventing access to or observation of the interior. At test time, after the agent had reached the goal for the first time, all doors were opened. All the agents were tested for 100 episodes, each one lasting for a fixed duration of 5,400 environment steps (90 s).

Agent architectures. *Architecture for the grid cell agent.* The agent architecture (Extended Data Fig. 5) was composed of a visual module, the grid cell network (described above), and an actor–critic learner[41]. The visual module was a neural network with input consisting of a three-channel (RGB) 64 × 64 image $\phi \in [-1, 1]^{3\times84\times84}$. The image was processed by a convolutional neural network (see Supplementary Methods for details), which produced embeddings, $\vec{e}$, which in turn were used as input to a fully connected linear layer trained in a supervised fashion to predict place and head-direction cell ensemble activations, $\vec{c}$ and $\vec{h}$ (as specified above), respectively. The predicted place and head direction cell activity patterns were provided as input to the grid network 5% of the time on average, akin to occasional imperfect observations made by behaving animals of salient environmental cues[27]. Specifically, the output of the convolutional network $\vec{e}$ was then passed through a masking layer which zeroed the units with a probability of 95%.

The grid cell network of the agent was implemented as in the supervised learning set up except that the LSTM ('grid LSTM') was not initialized with ground truth place cell activations but instead set to zero. The inputs to the grid cell network were the two translation velocities, u and v, as in DeepMind Laboratory it is possible to move in a direction different from the facing direction; the sine and cosine of the angular velocity, $\mathring{\varphi}$, (these velocities are provided by DeepMind Laboratory); and the $\vec{y}$ and $\vec{z}$ output by the vision module. In contrast to the supervised learning case, here the grid cell network had to use $\vec{y}$ and $\vec{z}$ to learn how to reset its internal state each time it was teleported to an arbitrary location in the environment (for example, after visiting the goal). As in the supervised learning experiments described above, the configuration of place fields (that is, location of place field centres in the 11 × 11 environments, goal-driven and goal-doors, 10 × 10 square arena, and 13 × 13 double E-maze) were constant throughout training (that is, across episodes).

For the actor–critic learner, the input was a three-channel 64 × 64 image $\phi_t \in [-1, 1]^{3\times84\times84}$, which was processed by a convolutional neural network followed by a fully connected layer (see Supplementary Methods for details). The output of the fully connected layer of the convolutional network $\vec{e'}$ was then concatenated with the reward r_t, the previous action a_{t-1}, the current grid code $\vec{g}_t$, and the goal grid code $\vec{g}_*$ (that is, linear layer activations observed last time the goal was reached)—or zeros if the goal had not yet been reached in the episode. Note we refer to these linear layer activations as 'grid codes' for brevity, even though units in this layer also comprise units resembling head direction cells and border cells (Extended Data Fig. 6a). This concatenated input was provided to an LSTM with 256 units. The LSTM had two different outputs. The first output, the actor, is a linear layer with six units followed by a softmax activation function, which represents a categorical distribution over the agent's next action. The second output, the critic, is a single linear unit that estimates the value function. Note that we refer to this as the 'policy LSTM' for brevity, even though it also outputs the value function.

Comparison agents. We compared the performance of the grid cell agent against two agents specifically because they use a different representational scheme for space (that is, place cell agent, place cell prediction agent), and relate to theoretical models of goal-directed navigation from the neuroscience literature (for example[42,43]). We also compared the grid cell agent against a baseline deep reinforcement learning agent, Asynchronous Advantage Actor–Critic (A3C)[41].

Place cell agent. The place cell agent architecture is shown in Extended Data Fig. 8b, and described in more detail in Supplementary Methods. In contrast to the grid cell agent, the place cell agent used ground truth information: specifically, the ground-truth place, $\vec{c}_t$, and head-direction, $\vec{h}_t$, cell activations (as described above). These activity vectors were provided as input to the policy LSTM in an analogous way to the provision of grid codes in the grid cell agent.

Specifically, the output of the fully connected layer of the convolutional network $\vec{e}_t$ was concatenated with the reward r_t, the previous action a_{t-1}, the ground-truth current place code, $\vec{c}_t$, and the current head-direction code, $\vec{h}_t$, together with the ground truth goal place code, $\vec{c}_*$, and ground truth head direction code, $\vec{h}_*$, observed last time the goal was reached—or with zeros if the goal had not yet been reached in the episode (Extended Data Fig. 8b). The convolutional network had the same architecture as described for the grid cell agent.

Place cell prediction agent. The architecture of the place cell prediction agent (Extended Data Fig. 9a) is similar to that of the grid cell agent described above: the key difference is the nature of the input provided to the policy LSTM as described below. The place cell prediction agent had a grid cell network with the same parameters as that of the grid cell agent. However, instead of using grid codes from the linear layer of the grid network $\vec{g}$ as input for the policy LSTM (as in the grid cell agent), we used the predicted place cell population activity vector $\vec{y}$ and the predicted head direction population activity vector $\vec{z}$ (the activations present on the output place and head direction unit layers of the grid cell network at each timestep; see Supplementary Methods).

The critical difference between the place cell agent and the place cell prediction agent (see Extended Data Figs. 8b and 9a, respectively) is that the former used ground truth information (place and head direction cell activations for current location and goal location), whereas the latter used the population activity produced across the output place and head direction cell layers (for current location and goal location) by the linear layer of the same grid network as used by the grid cell agent.

A3C. We implemented the asynchronous advantage actor–critic architecture[41] with a convolutional network having the same architecture as described for the grid cell agent (Extended Data Fig. 8a).

Other agents. We also assessed the performance of two deep reinforcement learning agents with external memory (Extended Data Fig. 9b), which served to establish the challenging nature of the multi-compartment environments (goal-doors and goal-driven). First, we implemented a memory network agent (NavMemNet) consisting of the FRMQN architecture[3] but instead of Q-learning we used the A3C algorithm described below. Furthermore, the input to memory was generated as an output from the LSTM controller (Extended Data Fig. 9b), rather than constituting embeddings from the convolutional network (that is, as in ref. 3). The convolutional network had the same architecture as described for the grid cell agent and the memory was formed of two banks (keys and values), each one with 1,350 slots.

Second, we implemented a differentiable neural computer (DNC) agent that uses content-based retrieval and writes to the most recently used or least recently used memory slot[44].

Training algorithms. We used the A3C algorithm[41], which implements a policy $\pi\,(a|s, \theta)$ and an approximation to its value function $V\,(s, \theta)$ using a neural network parameterised by θ. A3C adjusts the network parameters using n-step lookahead values, $\hat{R}_t = \sum_{i=0\ldots n-1} \gamma^i r_{t+1} + \gamma^n V\,(s_{t+n}, \theta)$, to minimize: $\mathcal{L}_{A3C} = \mathcal{L}_\pi + \alpha\mathcal{L}_V + \beta\mathcal{L}_H$, where $\mathcal{L}_\pi = -\mathbb{E}_{s_t\sim\pi}[\hat{R}_t]$, $\mathcal{L}_V = -\mathbb{E}_{s_t\sim\pi}[(\hat{R}_t - V\,(s_t, \theta))^2]$, $\mathcal{L}_\pi = -\mathbb{E}_{s_t\sim\pi}[H\,(\pi(\cdot|\, s_t, \theta))]$ where $\mathcal{L}_H$ is a policy entropy regularization term (see Supplementary Methods for details of the reinforcement learning approach). The grid cell network and the vision module were trained with the same loss reported for the supervised learning: $\mathcal{L}(\vec{y}, \vec{z}, \vec{c}, \vec{h}) = -\sum_{i=1}^N c_i \log(y_i) - \sum_{j=1}^N h_j \log(z_j)$.

Agent training details. We followed closely a previously described approach[41]. Each experiment used 32 actor–critic learner threads running on a single CPU machine. All threads applied updates to their gradients every four actions (that is, action repeat of 4) using RMSProp with shared gradient statistics[41]. All the experiments were run for a total of 10^9 environment steps.

In architectures where the grid cell network and the vision module were present, we used a shared buffer[45,46] in which we stored the agent's experiences at each time-step, $e_t = (\phi_t, u_t, v_t)$, collected over many episodes. All the 32 actor–critic workers were updating the same shared buffer which had a total size of 20×10^6. The vision module was trained with mini batches of size 32 frames ($\vec{\phi}$) sampled randomly from the replay buffer. The grid cell network was trained with mini batches of size 10, randomly sampled from the buffer, each one comprising a sequence of 100 consecutive observations, $[\vec{\phi}, \vec{u}, \vec{v}]$. These mini batches were first forwarded through the vision module to get $\vec{c}$, and $\vec{h}$, which were then passed through a masking layer that masked them to 0 with a probability of 95% (as described above in section on grid cell architecture). The output of this masking layer was then concatenated with $\vec{u}$, $\vec{v}$, $\overrightarrow{sin\mathring{\varphi}}$, $\overrightarrow{cos\mathring{\varphi}}$, which were then used as inputs to the grid network, as previously described (see Extended Data Fig. 5 for details). Both networks were trained using one single thread, one to train the vision module and another to train the grid network (so in total we used 34 threads). Also, there was no gradient sharing between the actor–critic learners, the vision module and the grid network.

The hyperparameters of the grid cell network were kept fixed across all the simulations and were derived from the best performing network in the supervised learning experiments. For the hyperparameter details of the vision module, the grid network and the actor–critic learner, see Supplementary Table 2 in Supplementary Methods. For each of the agents in this paper, 60 replicas were run with hyperparameters sampled from the same interval and different initial random seeds.

Details of lesion experiment. To conduct a lesioning experiment in the agent, we trained the grid cell agent with dropout applied on the goal grid code input $\vec{g}_t$. Specifically, every 100 training steps we generated a random mask to silence 20% of the units in the goal grid code ($\vec{g}_*$)—that is, units were zeroed. This procedure was implemented to ensure that the policy LSTM would become robust through training to receiving a lesioned input (that is, would not catastrophically fail), and still be able to perform the task.

We then selected the agent with the best performance over 100 episodes, and we computed the grid score of all units found in $\vec{g}$. The critical comparison to test the importance of grid-like units to vector-based navigation was as follows. In one condition we ran 100 testing episodes where we silenced the 25% units in $\vec{g}_*$ with the highest grid scores. In the other condition, we ran 100 testing episodes with the same agent with 25% random units in $\vec{g}_*$ silenced. In this second case, we ensured that head direction cells with a resultant vector length of more than 0.47 were not silenced, to preserve crucial head direction signals. We then compared the performance, and the representation of metrics relating to vector-based navigation, of the agents under these two conditions.

Details of experiment using 'fake' goal grid code. To demonstrate that the goal grid code provided sufficient information to enable the agent to navigate to an arbitrary location, we took an agent trained in the square arena, froze the weights, and ran it in the same square arena for 5,400 steps. Critically, after the sixth time the agent reached the goal, we sampled the grid code from a random point that the agent visited in the environment (the fake goal grid code). We then substituted the true goal grid code with this fake goal grid code, to show that this would be sufficient to direct the agent to a location where there was no actual goal.

Agent performance. To evaluate agent performance during training (as in Figs. 2f, 3e, f) we selected the 30 replicas (out of 60) with the highest average cumulative reward across 100 episodes. We also assessed the robustness of the architecture over different initial random seeds and the hyperparameters in Supplementary Methods Table 2 by calculating the area under the curve (AUC). To plot the AUC we ran 60 replicas with hyperparameters sampled from the same interval (see Supplementary Table 2 in Supplementary Methods) and different initial random seeds (Extended Data Fig. 7a–c).

Neuroscience-based analyses of network units. *Generation of activity maps.* Spatial (ratemaps) and directional activity maps were calculated for individual units as follows. Each point in the trajectory was assigned to a specific spatial and directional bin according to its location and the direction in which it faced. Spatial bins were defined as a 32×32 square grid spanning each environment and directional bins as 20 equal width intervals. Then, for each unit, the mean activity over all the trajectories points assigned to that bin was found. These values were displayed and analysed further without additional smoothing.

Inter-trial stability. For each unit, the reliability of spatial firing between baseline trials was assessed by calculating the spatial correlation between pairs of rate maps taken at two different logging steps in training ($t = 2 \times 10^5$; $t' = 3 \times 10^5$). The total training time was 3×10^5, so the points were selected with enough time difference to minimize the chances of finding random correlations. The Pearson product moment correlation coefficient was calculated between equivalent bins in the two trials and unvisited bins were excluded from the measure.

Quantification of spatial activity. Where possible, we assessed the spatial modulation of units using measures adopted from the neuroscience literature. The hexagonal regularity and scale of grid-like patterns were quantified using the gridness score[18,20] and grid scale[20], measures derived from the spatial autocorellogram[20] of each unit's ratemap. Similarly, the degree of directional modulation exhibited by each unit was assessed using the length of the resultant vector[47] of the directional activity map. Finally, the propensity of units to fire along the boundaries of the environment was quantified using the border score[48].

The gridness and border scores exhibited by units in the linear layer were benchmarked against the 95th percentile of null distributions obtained using a permutation procedure (spatial field shuffle[49]) applied to each unit's ratemap. This shuffling procedure aimed to preserve the local topography of fields within each ratemap while distributing the fields themselves at random[49]. The means, over units, of the thresholds obtained were gridness >0.37 and border score >0.50. Units exceeding these thresholds were considered to be grid-like and border-like, respectively. To identify directionally modulated cells, we applied Rayleigh tests of directional uniformity to the binned directional activity maps. A unit was considered to be directionally modulated if the null hypothesis of uniform was rejected at the $\alpha = 0.01$ level, corresponding to units with resultant vector length in excess of 0.47 (see Supplementary Methods).

Clustering of scale in grid-like units. To determine whether grid-like units exhibited a tendency to cluster around specific scales, we applied two methods. First, to determine whether the scales of grid-like units (gridness > 0.37, 129/512 units) followed a continuous or discrete distribution, we calculated the discreteness measure[22] of the distribution of their scales (see Supplementary Methods). The discreteness score of the real data was found to exceed that of all of the 500 shuffles. Second, to characterize the number and location of scale clusters, the distribution of scales from grid-like units was fit with Gaussian mixture distributions, and three components were found to provide the most parsimonious fit, indicating the presence of three scale clusters (See Supplementary Methods).

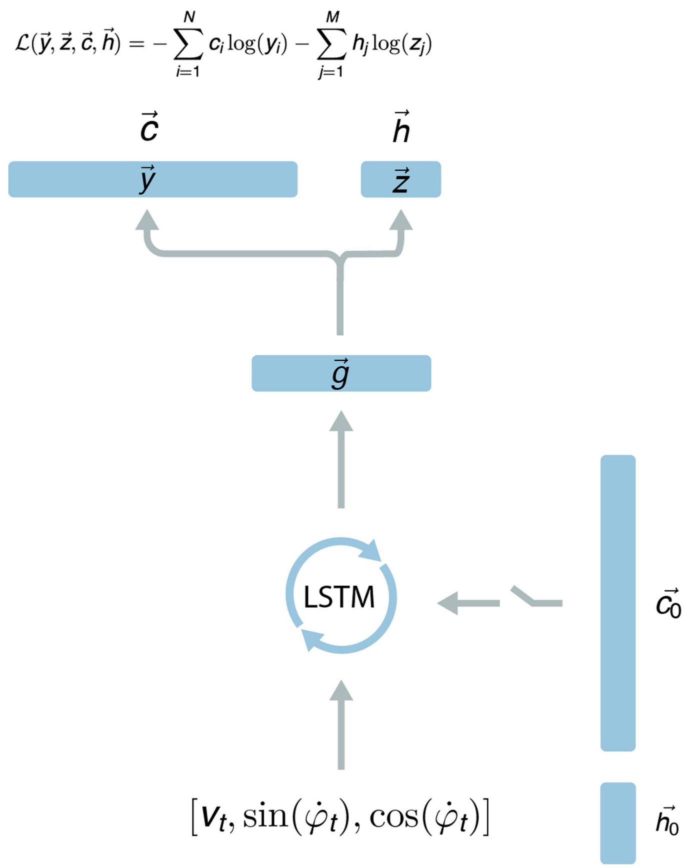
Multivariate decoding of representation of metric quantities within LSTM. To test whether the grid agent learns to use the predicted vector based navigation (VBN) metric codes, we recorded the activation from the hidden units of the policy LSTM layer while the agent navigated 200 episodes in the land maze. We used L2-regularized (ridge) regression to decode Euclidean distance and allocentric direction to the goal (see Supplementary Methods for full decoding details). We specifically focused on twelve steps (steps 9–21) during the early portion of navigation, but after the agent has had time to accurately self-localize. It is this early period after the agent has reached the goal for the first time in which a VBN strategy should be most effective. We conducted the same analysis on the place cell agent control, which is not predicted to use vector-based navigation as efficiently. The decoding accuracy was measured as the correlation between predicted and actual metric values in held-out data. Decoding accuracy was compared across different agents by assessing the difference in decoding correlations between the agents. A bootstrap method (using 10,000 samples) was used to compute a 95% confidence interval on this correlation difference, and these are reported for each comparison. The same approach was used to decode and compare these two metrics in the lesioned grid agents on the land maze. Finally, to explore VBN metrics in a more complex environment, the same method was applied to the goal-driven task. In this case we also investigated metric decoding in the control A3C agent.

Reporting summary. Further information on experimental design is available in the Nature Research Reporting Summary linked to this paper.

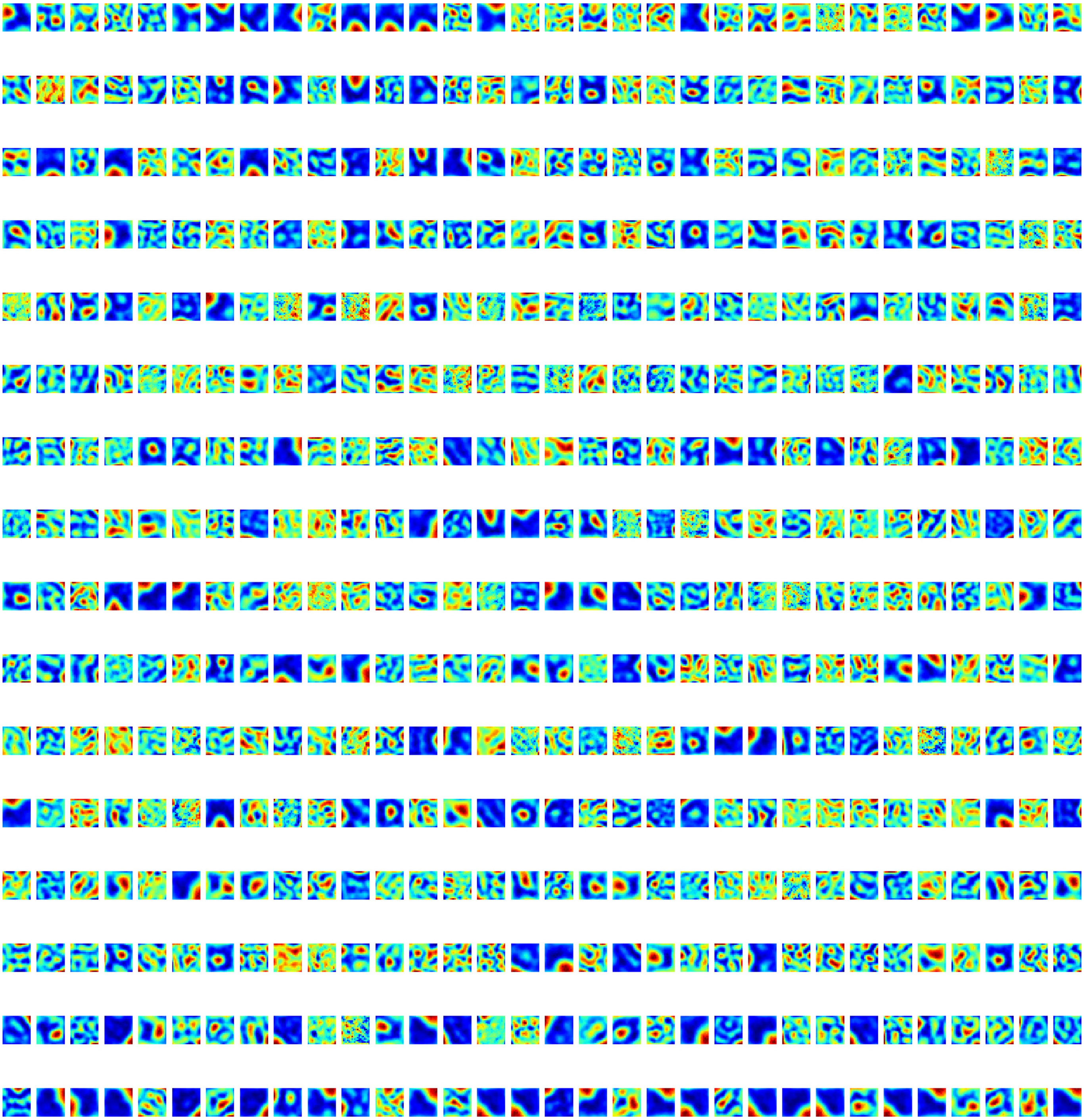
Data availability statement. All reinforcement learning tasks described throughout the paper were built using the publicly available DeepMind Laboratory platform (https://github.com/deepmind/lab). Both the goal driven and goal doors tasks are included as part of the latest release, named explore_goal_locations, and explore_obstructed_goals, respectively.

Code availability statement. We will release the code for the supervised learning experiments within the next six months. The codebase for the deep reinforcement learning agents makes use of proprietary components, and we are unable to publicly release this code. However, all experiments and agents are described in sufficient detail to allow independent replication.

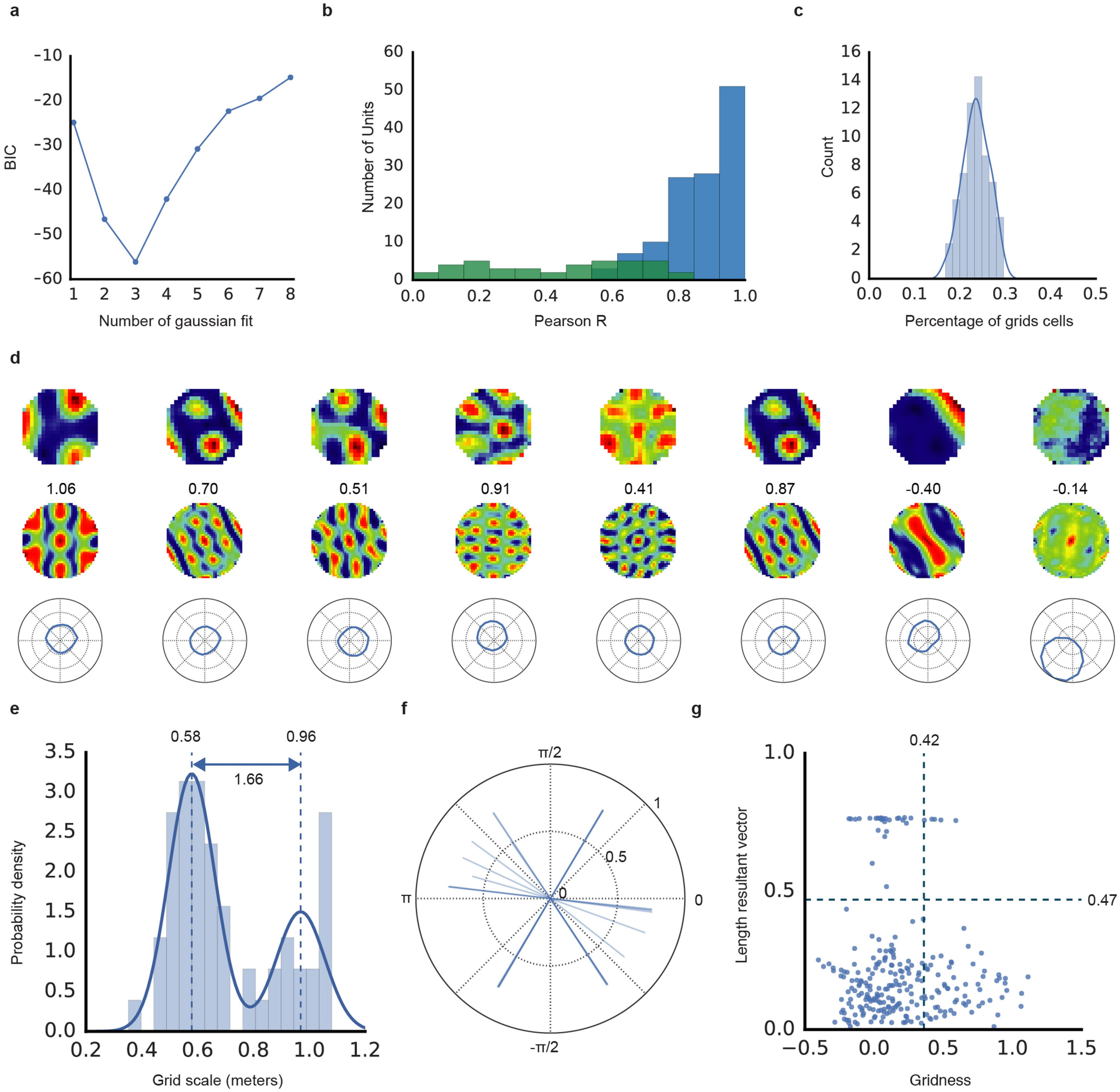
31. Raudies, F. & Hasselmo, M. E. Modeling boundary vector cell firing given optic flow as a cue. *PLOS Comput. Biol.* **8**, e1002553 (2012).
32. Hochreiter, S. & Schmidhuber, J. Long short-term memory. *Neural Comput.* **9**, 1735–1780 (1997).
33. Bridle, J. S. in Touretzky, D. S. (ed.) *Advances in Neural Information Processing Systems 2* 211–217 (Morgan-Kaufmann, 1990).
34. Elman, J. L. & McClelland, J. L. Exploiting lawful variability in the speech wave. *Invariance and Variability in Speech Processes* **1**, 360–380 (1986).
35. Tieleman, T. & Hinton, G. Lecture 6.5—RmsProp: Divide the gradient by a running average of its recent magnitude. COURSERA: Neural Networks for Machine Learning (2012).
36. MacKay, D. J. A practical bayesian framework for backpropagation networks. *Neural Comput.* **4**, 448–472 (1992).
37. Pascanu, R., Mikolov, T. & Bengio, Y. On the difficulty of training recurrent neural networks. Proc. 30th ICML 28, 1310–1318 (2013).
38. Ackley, D. H., Hinton, G. E. & Sejnowski, T. J. A learning algorithm for Boltzmann machines. *Cogn. Sci.* **9**, 147–169 (1985).
39. Beattie, C. et al. Deepmind lab. Preprint at https://arxiv.org/abs/1612.03801 (2016).
40. Doeller, C. F., Barry, C. & Burgess, N. Evidence for grid cells in a human memory network. *Nature* **463**, 657–661 (2010).
41. Mnih, V. et al. Asynchronous methods for deep reinforcement learning. In *Proc. 33nd Intl Conf. Machine Learning* 1928–1937 (2016).
42. Touretzky, D. S. & Redish, A. D. Theory of rodent navigation based on interacting representations of space. *Hippocampus* **6**, 247–270 (1996).
43. Foster, D. J., Morris, R. G. & Dayan, P. A model of hippocampally dependent navigation, using the temporal difference learning rule. *Hippocampus* **10**, 1–16 (2000).
44. Graves, A. et al. Hybrid computing using a neural network with dynamic external memory. *Nature* **538**, 471–476 (2016).
45. Mnih, V. et al. Human-level control through deep reinforcement learning. *Nature* **518**, 529–533 (2015).
46. Lin, L.-J. Reinforcement learning for robots using neural networks. Technical Report (Carnegie-Mellon Univ. School of Computer Science, 1993).
47. Knight, R. et al. Weighted cue integration in the rodent head direction system. *Phil. Trans. R. Soc. Lond. B* **369**, 20120512 (2013).
48. Solstad, T., Boccara, C. N., Kropff, E., Moser, M.-B. & Moser, E. I. Representation of geometric borders in the entorhinal cortex. *Science* **322**, 1865–1868 (2008).
49. Barry, C. & Burgess, N. To be a grid cell: Shuffling procedures for determining gridness. Preprint at https://www.biorxiv.org/content/early/2017/12/08/230250 (2017).



Extended Data Fig. 1 | Network architecture in the supervised learning experiment. The recurrent layer of the grid cell network is an LSTM with 128 hidden units. The recurrent layer receives as input the vector $[\vec{v}, \sin(\mathring{\varphi}), \cos(\mathring{\varphi})]$. The initial cell state and hidden state of the LSTM, $\vec{l}_0$ and $\vec{m}_0$, respectively, are initialized by computing a linear transformation of the ground truth place $\vec{c}_0$ and head-direction activity $\vec{h}_0$ at time 0. The output of the LSTM is followed by a linear layer on which dropout is applied. The output of the linear layer, $\vec{g}_t$, is linearly transformed and passed to two softmax functions that calculate the predicted head direction cell activity, $\vec{z}_t$, and place cell activity, $\vec{y}_t$. We found evidence of grid-like and head direction-like units in the linear layer activations $\vec{g}_t$.

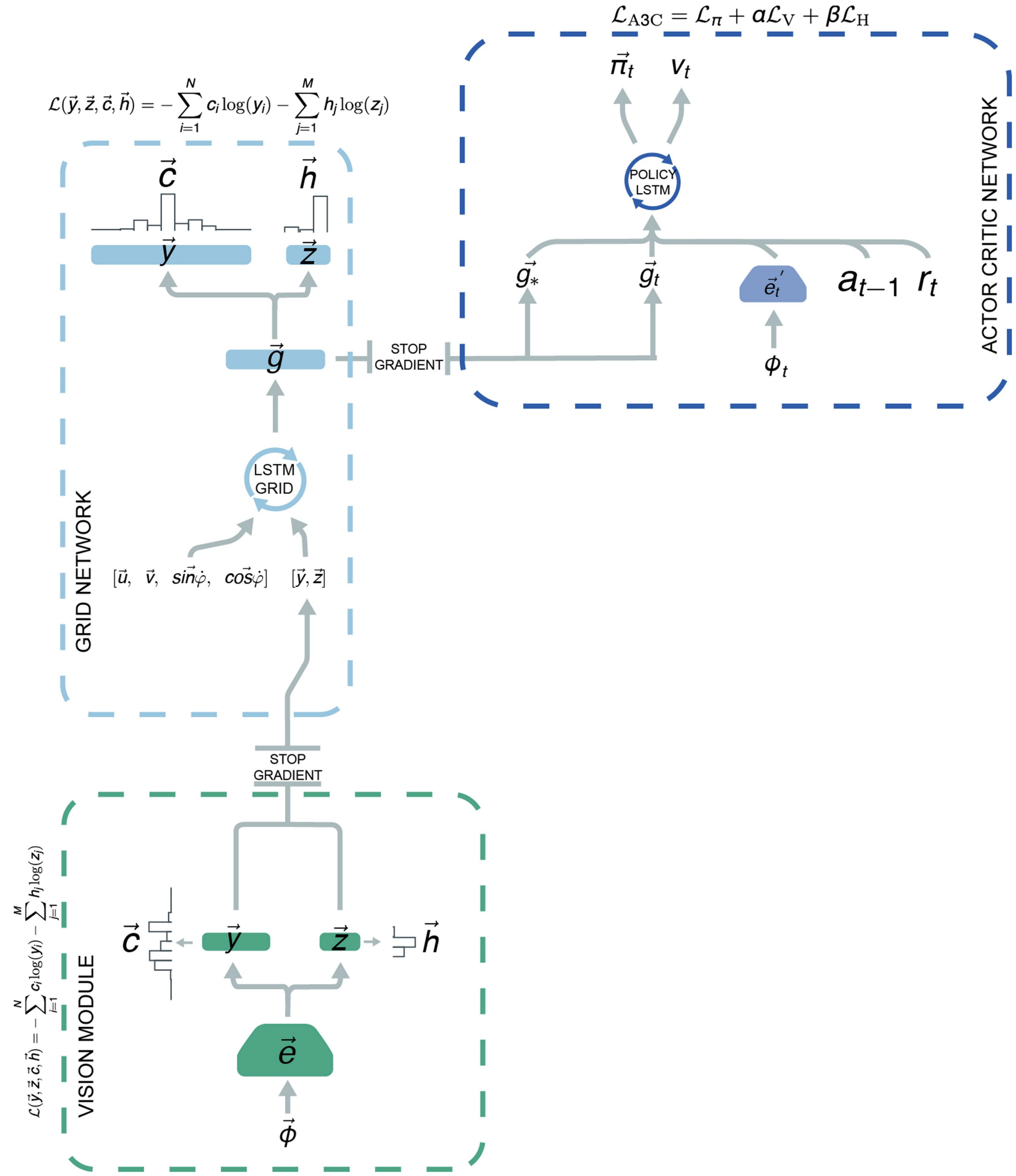


Extended Data Fig. 2 | Linear layer spatial activity maps from the supervised learning experiment. Spatial activity plots for all 512 units in the linear layer $\vec{g}_t$. Units exhibit spatial activity patterns resembling grid cells, border cells, and place cells. Head direction tuning was also present but is not shown.

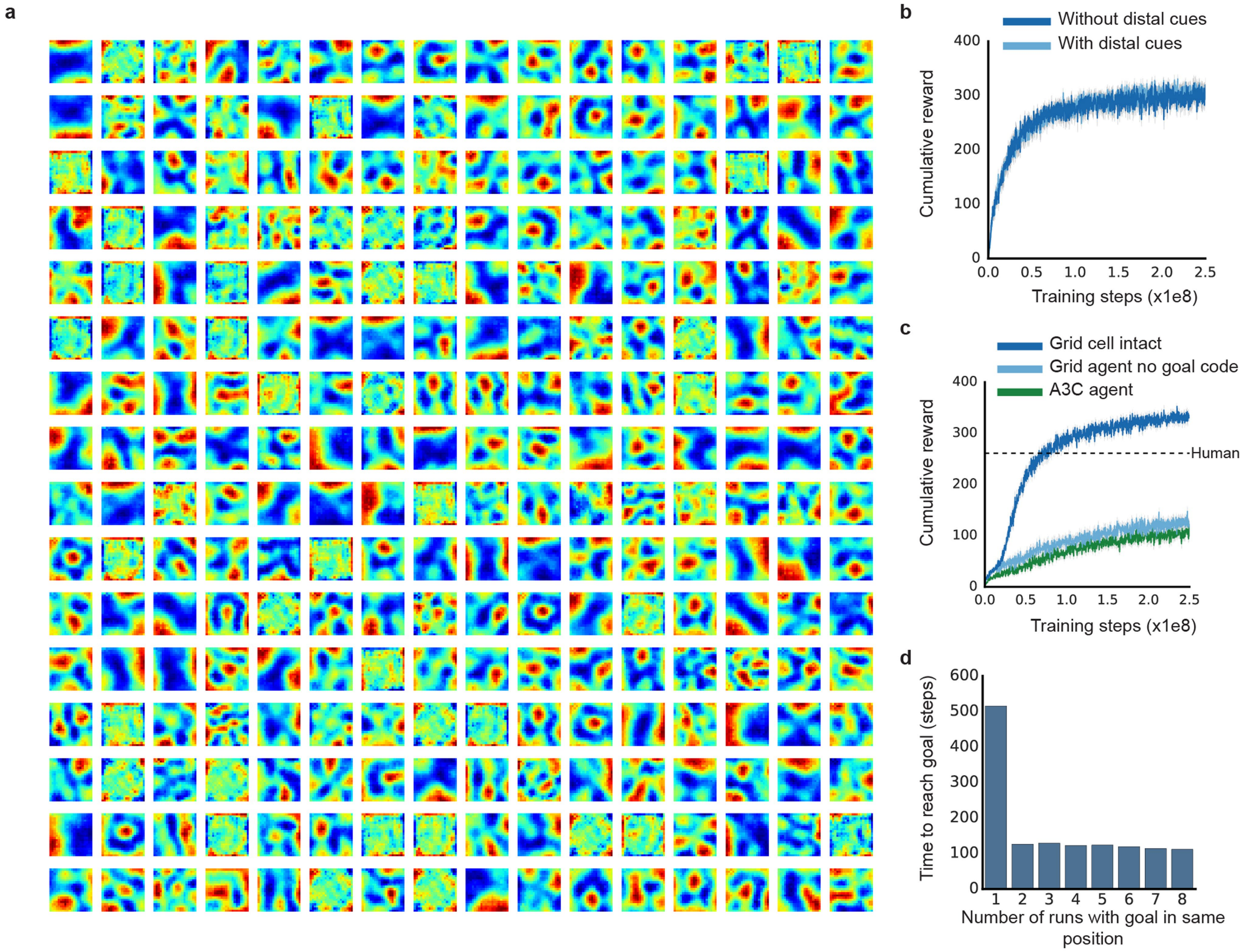


Extended Data Fig. 3 | Characterization of grid-like units in square environment and circular environment. a, The scale (assessed from the spatial autocorrelogram of the ratemaps) of grid-like units exhibited a tendency to cluster at specific values. The number of distinct scale clusters was assessed by sequentially fitting Gaussian mixture models with one to eight components. In each case, the efficiency of the fit (likelihood versus number of parameters) was assessed using Bayesian information criterion (BIC). BIC was minimized with three Gaussian components, indicating the presence of three distinct scale clusters. **b**, Spatial stability of units in the linear layer of the supervised network was assessed using spatial correlations— bin-wise Pearson product moment correlation between spatial activity maps (32 spatial bins in each map) generated at two different points in training, $t = 2 \times 10^5$ and $t' = 3 \times 10^5$ training steps (two-thirds of the way through training and at the end of training, respectively). This separation was imposed to minimize the effect of temporal correlations and to provide a conservative test of stability. Grid-like units (gridness > 0.37), blue; directionally modulated units (resultant vector length > 0.47, green. Grid-like units exhibit high spatial stability, while directionally modulated units do not. **c**, Robustness of the grid representation to starting conditions. The network was retrained 100 times with the same hyperparameters but different random seeds controlling the initialization of network weights, $\overrightarrow{c}$ and $\overrightarrow{h}$. Populations of grid-like units (gridness > 0.37) were found to appear in all cases, with the average proportion of grid-like units being 23% (s.d. 2.8%). **d**, The supervised network was also trained in a circular environment (diameter 2.2 m). As before, units in the linear layer exhibited spatially tuned responses resembling grid, border, and head direction cells. Eight units are shown. Top, ratemap displaying activity binned over location. Middle, spatial autocorrelogram of the ratemap; gridness[20] is indicated above. Bottom, polar plot of activity binned over head direction. **e**, Spatial scale of grid-like units ($n = 56$ (21.9%)) is clustered. Distribution is best fit by a mixture of two Gaussians (centres 0.58 and 0.96 m, ratio 1.66). **f**, Distribution of directional tuning for 31 most directionally active units; single line for each unit indicates length and orientation of resultant vector[47]. **g**, Distribution of gridness and directional tuning. Dashed lines indicate 95% confidence interval derived from shuffling procedure (500 permutations); five grid units (9%) exhibit significant directional modulation.

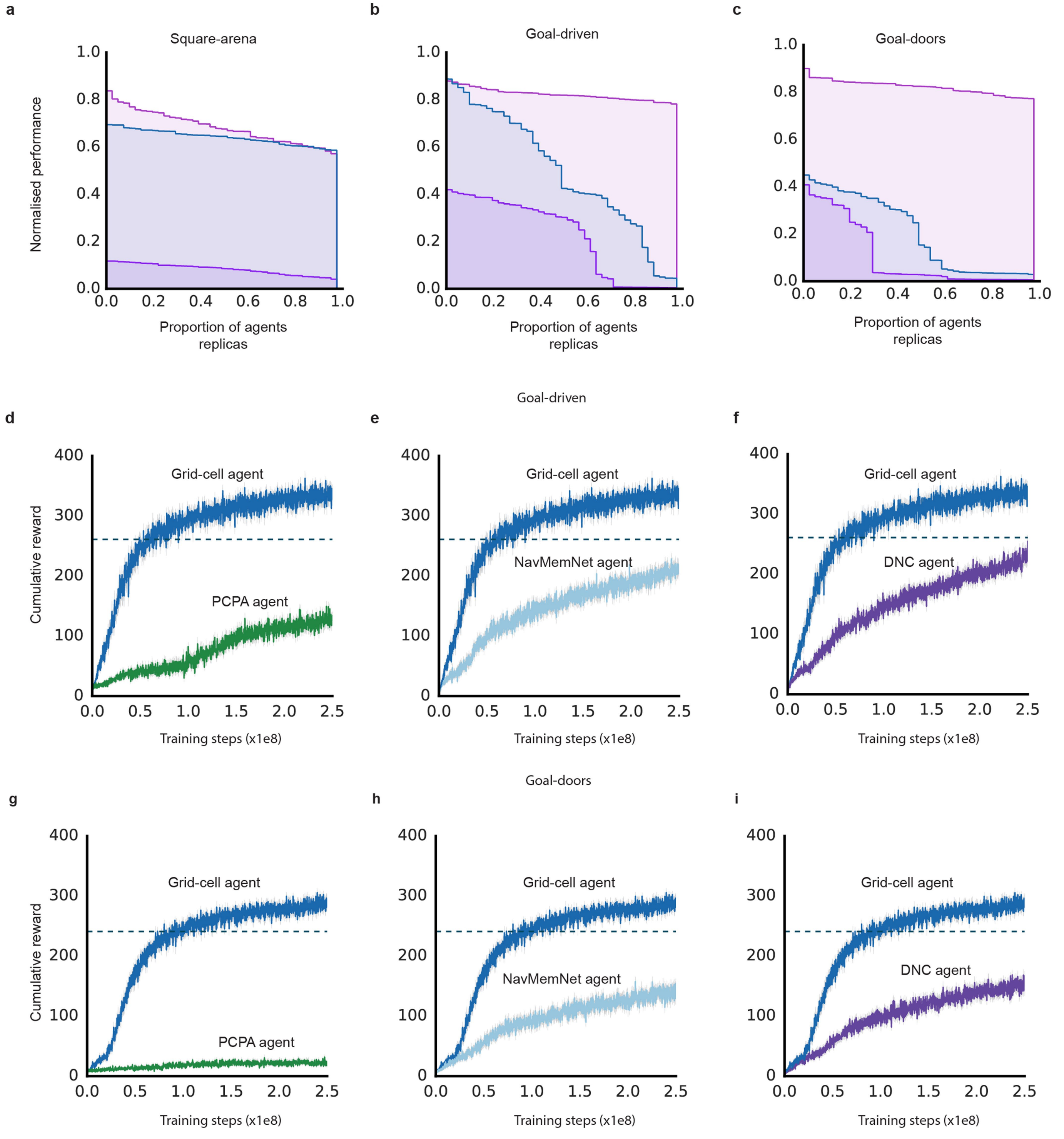
Extended Data Fig. 4 | Grid-like units did not emerge in the linear layer when dropout was not applied. Linear layer spatial activity maps ($n = 512$) generated from a supervised network trained without dropout. The maps do not exhibit the regular periodic structure diagnostic of grid cells.



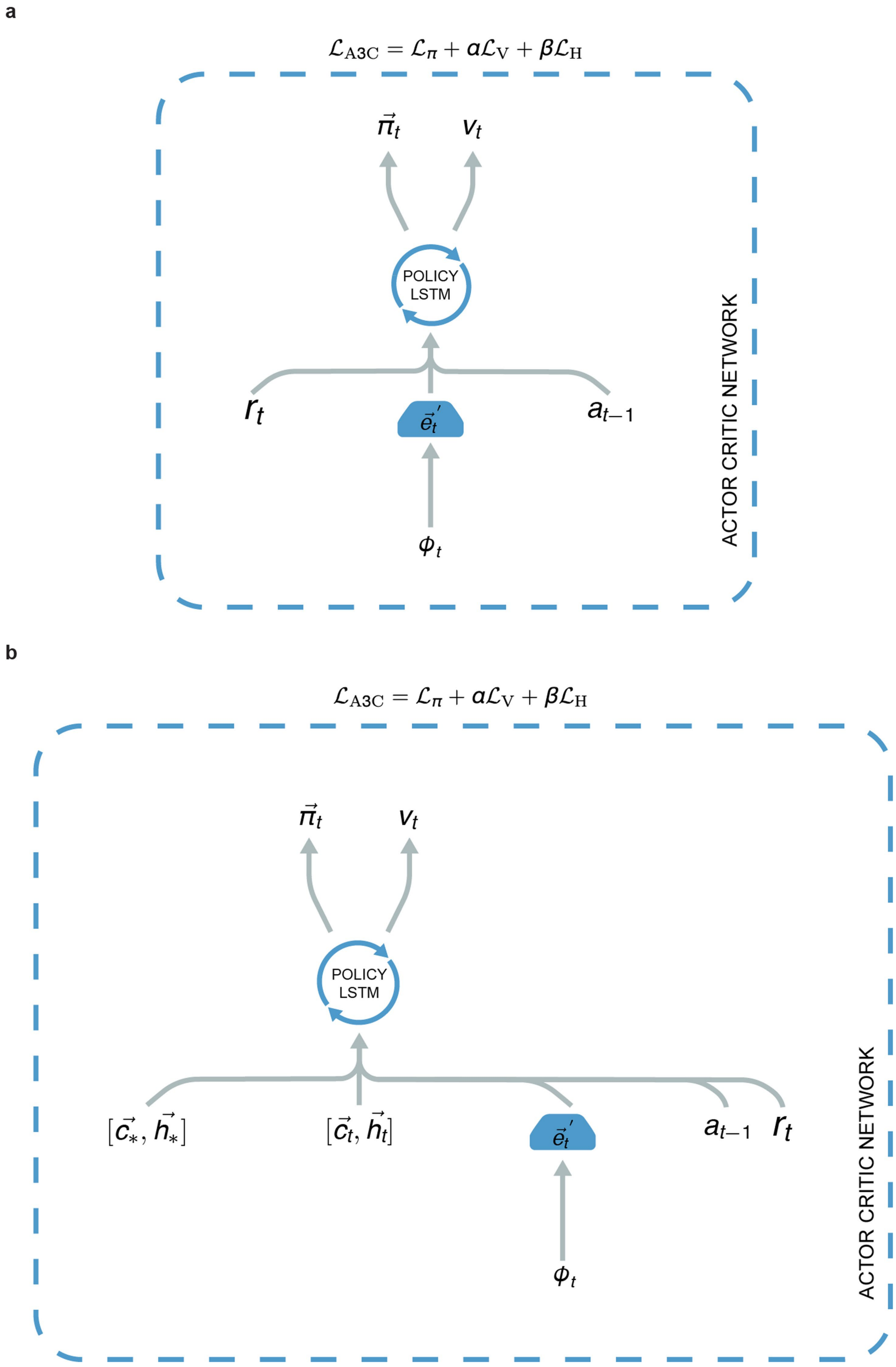
Extended Data Fig. 5 | Architecture of the grid cell agent. The architecture of the supervised network (grid network, light blue dashed) was incorporated into a larger deep reinforcement learning network, including a visual module (green dashed) and an actor–critic learner (based on A3C[41]; dark blue dashed). In this case the supervised learner does not receive the ground truth $\vec{c}_0$ and $\vec{h}_0$ to signal its initial position, but uses input from the visual module to self-localize after placement at a random position within the environment. Visual module: since experimental evidence suggests that place cell input to grid cells functions to correct for drift and anchor grids to environmental cues[21,27], visual input was processed by a convolutional network to produce place cell (and head direction cell) activity patterns which were used as input to the grid network. The output of the vision module was only provided 5% of the time to the grid network (see Methods for implementational details), akin to occasional observations of salient environmental cues made by behaving animals[27]. The output of the vision module was concatenated with $\vec{u}$, $\vec{v}$, $sin\dot{\vec{\varphi}}$, $cos\dot{\vec{\varphi}}$ to form the input to the grid LSTM, which is the same network as in the supervised case (see Methods and Extended Data Fig. 1). The actor–critic learner (light blue dashed) receives as input the concatenation of $\vec{e}_t{}'$ produced by a convolutional network with the reward r_t, the previous action a_{t-1}, the linear layer activations of the grid cell network $\vec{g}_t$ (current grid-code), and the linear layer activations observed last time the goal was reached, $\vec{g}_*$ (goal grid-code), which is set to zero if the goal has not been reached in the episode. The fully connected layer was followed by an LSTM with 256 units. The LSTM has two different outputs. The first output, the actor, is a linear layer with six units followed by a softmax activation function, which represents a categorical distribution over the agent's next action $\vec{\pi}_t$. The second output, the critic, is a single linear unit that estimates the value function v_t.



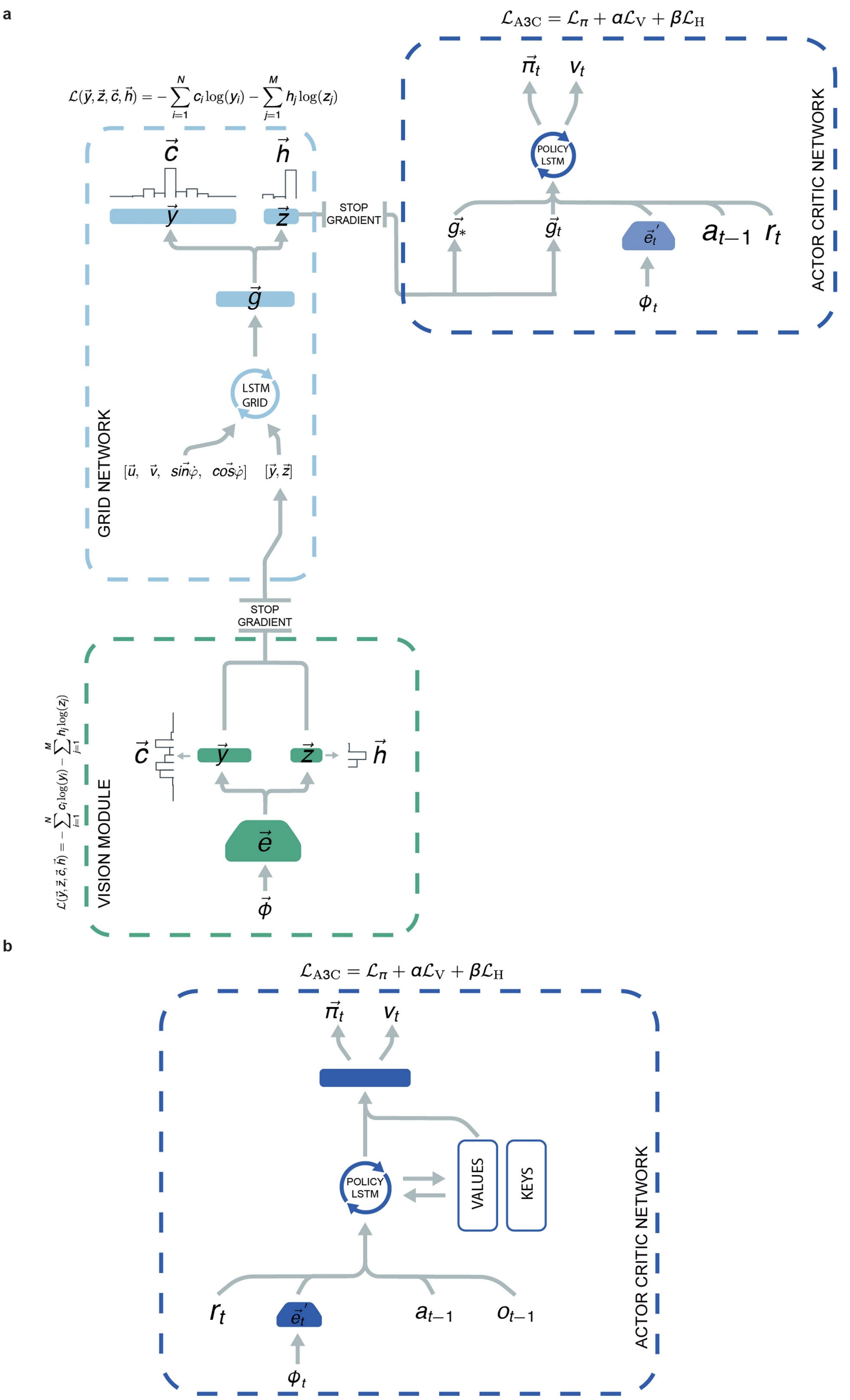
Extended Data Fig. 6 | Characterization of grid-like representations and robustness of performance for the grid cell agent in the square land maze environment. a, Spatial activity plots for the 256 linear layer units in the agent exhibit spatial patterns similar to grid, border, and place cells. **b**, Cumulative reward indexing goal visits per episode (goal, 10 points) when distal cues are removed (dark blue) and when distal cues are present (light blue). Performance is unaffected, hence dark blue largely obscures light blue. Average of 50% best agent replicas ($n = 32$) plotted (see Methods). The grey band displays the 68% CI based on 5,000 bootstrapped samples. **c**, Cumulative reward per episode when no goal code was provided (light blue) and when goal code was provided (dark blue). When no goal code was provided the agent performance fell to that of the baseline deep reinforcement learning agent (A3C) (100 episodes average score no goal code, 123.22 versus A3C, 112.06; effect size, 0.21; 95% CI, 0.18–0.28). Average of 50% best agent replicas ($n = 32$) plotted (see Methods). The grey band displays the 68% CI based on 5,000 bootstrapped samples. **d**, After locating the goal for the first time during an episode, the agent typically returned directly to it from each new starting position, showing decreased latencies for subsequent visits, paralleling the behaviour exhibited by rodents.



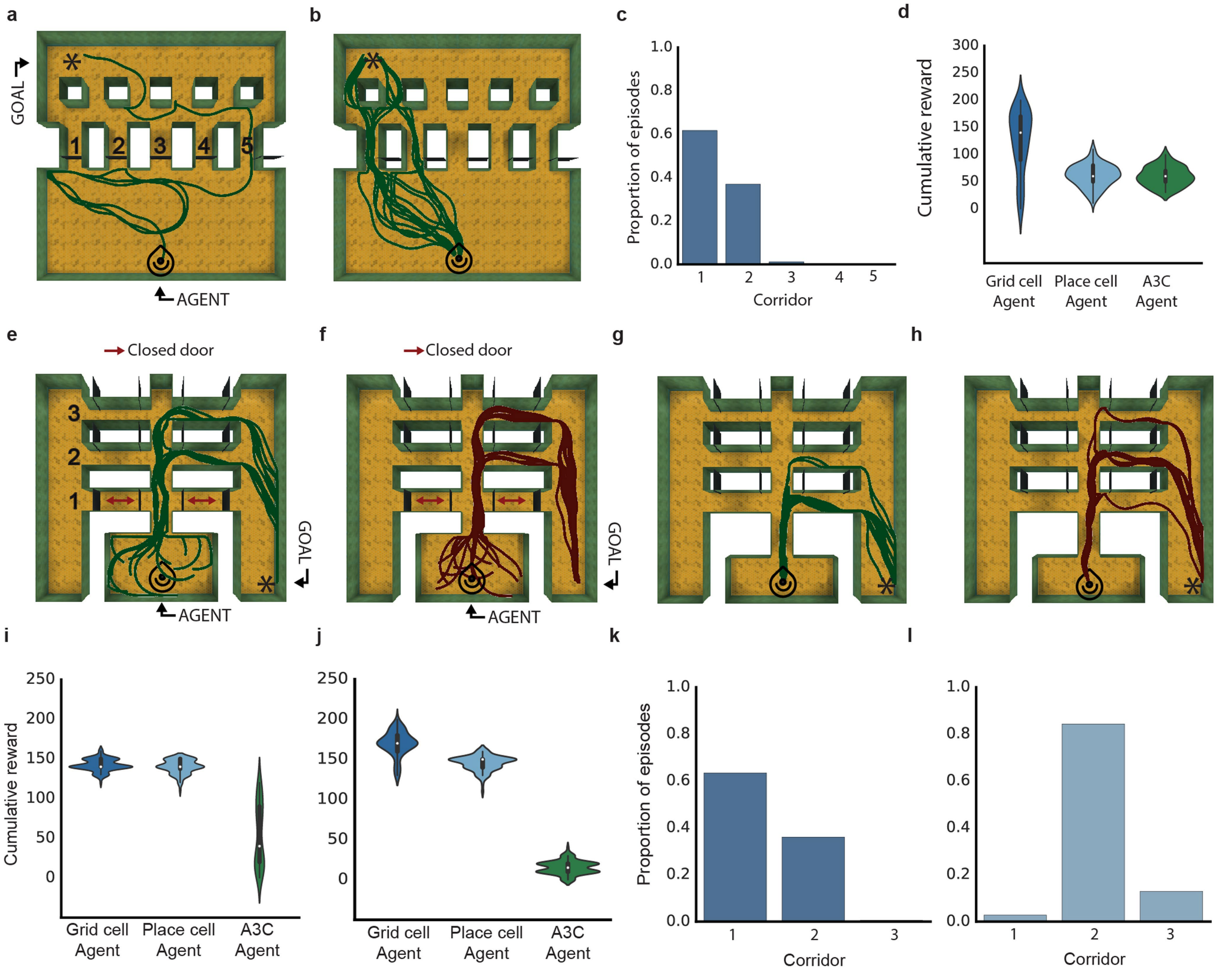
Extended Data Fig. 7 | Robustness of grid cell agent and performance of other agents. a–c, AUC performance gives robustness to hyperparameters (that is, learning rate, baseline cost, entropy cost; see Supplementary Table 2 in Supplementary Methods for details of the range) and seeds (see Methods). For each environment we run 60 agent replicas (see Methods). Light purple is the grid agent, blue is the place cell agent and dark purple is A3C. **a**, Square arena. **b**, Goal-driven. **c**, Goal doors. In all cases the grid cell agent shows higher robustness to variations in hyperparameters and seeds. **d–i**, Performance of place cell prediction, NavMemNet and DNC agents (see Methods) against grid cell agent. Dark blue is the grid cell agent (Extended Data Fig. 5), green is the place cell prediction agent (Extended Data Fig. 9a), purple is the DNC agent, light blue is the NavMemNet agent (Extended Data Fig. 9b). The grey band displays the 68% CI based on 5,000 bootstrapped samples. **d–f**, Performance in goal-driven. **g–i**, Performance in goal-doors. Note that the performance of the place cell agent (Extended Data Fig. 8b, lower panel) is shown in Fig. 3.



Extended Data Fig. 8 | Architecture of the A3C and place cell agent. **a**, The A3C implementation is as described[41]. **b**, The place cell agent was provided with the ground-truth place, $\vec{c}_t$, and head-direction, $\vec{h}_t$, cell activations (as described above) at each time step. The output of the fully connected layer of the convolutional network $\vec{e}_t$ was concatenated with the reward r_t, the previous action a_{t-1}, the ground-truth current place code, $\vec{c}_t$, and current head-direction code, $\vec{h}_t$, together with the ground truth goal place code, $\vec{c}_*$, and ground truth head direction code, $\vec{h}_*$, observed the last time the agent reached the goal (see Methods).



Extended Data Fig. 9 | Architecture of the place cell prediction agent and of the NavMemNet agent. a, The architecture of the place cell prediction agent is similar to the grid cell agent, having a grid cell network with the same parameters as that of the grid cell agent. The key difference is the nature of the input provided to the policy LSTM. Instead of using grid codes from the linear layer of the grid network $\vec{g}$, we used the predicted place cell population activity vector $\vec{y}$, and the predicted head direction population activity vector $\vec{z}$, (the activations present on the output place and head direction unit layers of the grid cell network, corresponding to the current and goal position, respectively) as input for the policy LSTM. As in the grid cell agent, the output of the fully connected layer of the convolutional network, $\vec{e}$, the reward r_t, and the previous action a_{t-1}, were also input to the policy LSTM. The convolutional network had the same architecture as described for the grid cell agent. **b**, NavMemNet agent. The architecture implemented is as described[3], specifically FRMQN, but the A3C algorithm was used in place of Q-learning. The convolutional network had the same architecture described for the grid cell agent and the memory was formed of two banks (keys and values), each composed of 1,350 slots.



Extended Data Fig. 10 | Flexible use of shortcuts. a, Overhead view of the linear sunburst maze in initial configuration, with only door 5 open. Example trajectory from grid cell agent during training (green line, icon indicates start location). **b**, Test configuration with all doors open; grid cell agent uses the newly available shortcuts (multiple episodes shown). **c**, Histogram showing proportion of times the agent uses each of the doors during 100 test episodes. The agent shows a clear preference for the shortest paths. **d**, Performance of grid cell agent and comparison agents during test episodes. **e**, **f**, Example grid cell agent (**e**) and example place cell agent (**f**) trajectory during training in the double E-maze (corridor 1 doors closed). **g**, **h**, In the test phase, with all doors open, the grid cell agent exploits the available shortcut (**g**), while the place cell agent does not (**h**). **i**, **j**, Performance of agents during training (**i**) and test (**j**). **k**, **l**, The proportion of times the grid (**k**) and place (**l**) cell agents used the doors on the first to third corridors during test. The grid cell agent shows a clear preference for available shortcuts, while the place cell agent does not.

LETTER

https://doi.org/10.1038/s41586-018-0092-4

An exclusive metabolic niche enables strain engraftment in the gut microbiota

Elizabeth Stanley Shepherd[1,2], William C. DeLoache[2], Kali M. Pruss[1], Weston R. Whitaker[2] & Justin L. Sonnenburg[1]*

The dense microbial ecosystem in the gut is intimately connected to numerous facets of human biology, and manipulation of the gut microbiota has broad implications for human health. In the absence of profound perturbation, the bacterial strains that reside within an individual are mostly stable over time[1]. By contrast, the fate of exogenous commensal and probiotic strains applied to an established microbiota is variable, generally unpredictable and greatly influenced by the background microbiota[2,3]. Therefore, analysis of the factors that govern strain engraftment and abundance is of critical importance to the emerging field of microbiome reprogramming. Here we generate an exclusive metabolic niche in mice via administration of a marine polysaccharide, porphyran, and an exogenous *Bacteroides* strain harbouring a rare gene cluster for porphyran utilization. Privileged nutrient access enables reliable engraftment of the exogenous strain at predictable abundances in mice harbouring diverse communities of gut microbes. This targeted dietary support is sufficient to overcome priority exclusion by an isogenic strain[4], and enables strain replacement. We demonstrate transfer of the 60-kb porphyran utilization locus into a naive strain of *Bacteroides*, and show finely tuned control of strain abundance in the mouse gut across multiple orders of magnitude by varying porphyran dosage. Finally, we show that this system enables the introduction of a new strain into the colonic crypt ecosystem. These data highlight the influence of nutrient availability in shaping microbiota membership, expand the ability to perform a broad spectrum of investigations in the context of a complex microbiota, and have implications for cell-based therapeutic strategies in the gut.

Changes to the microbial members of the highly competitive and dynamic gut microbiota can impact numerous aspects of host biology[5–7]. Despite the importance of gut microbe composition in human health, the rules governing invasion of commensal strains into an existing complex community are not well understood. Resident strains often appear to exclude similar invading strains[3,4] although in some cases, the opposite is true: the niche that is occupied by an existing strain can be exploited by a similar invading strain[8]. The inability to predict or control the compositional outcome of faecal microbiota transplants[9–11] illustrates the need for basic insight into the factors that influence whether new strains of bacteria can integrate into a pre-existing, complex microbiota.

To characterize the extent to which incoming exogenous bacteria variably colonize hosts with distinct microbiotas, we used a group of mice that had conventional mouse microbiota and two groups of

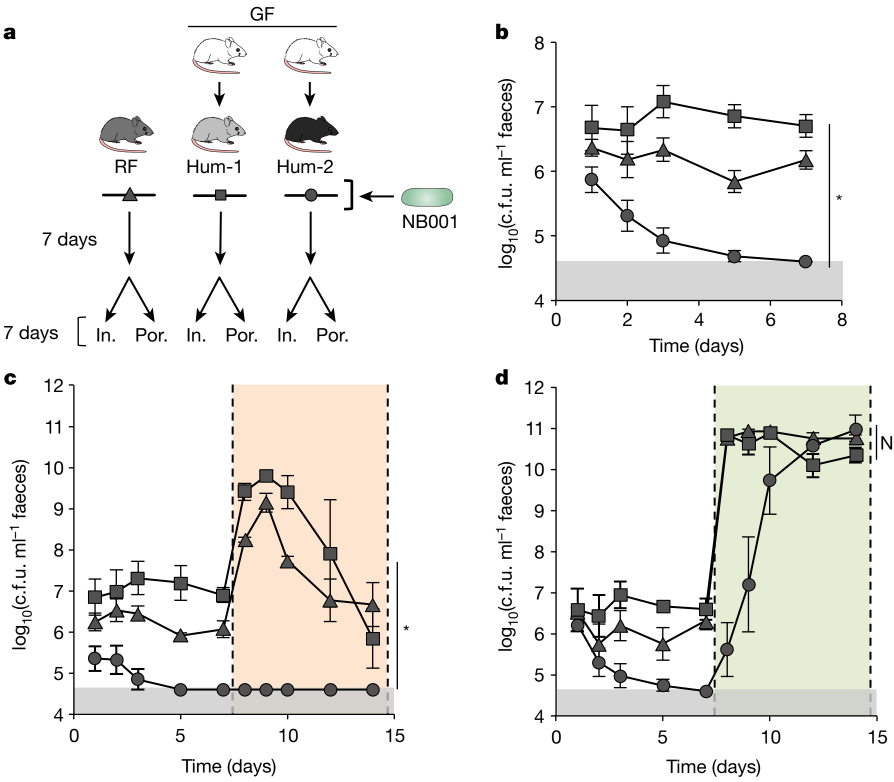


Fig. 1 | Niche availability varies by microbiota and can be modulated by addition of a privileged nutrient. a, Schematic of experimental design. Groups of mice with three different gut communities (equilibrated for four weeks on standard laboratory diet before seven-day exposure to MAC-deficient chow, see Methods) from mice (restricted flora (RF)) or humans (Hum-1, Hum-2), were colonized with NB001. NB001 was tracked in faeces for seven days, and mice were switched to specialized polysaccharide chow containing either inulin (in.) or porphyran-rich seaweed (por.). **b**, Density of NB001 in faeces in the three gut communities over the course of seven days. RF, $n = 9$; Hum-1, $n = 11$; Hum-2, $n = 10$; Kruskal–Wallis test, $^{*}P < 0.0001$. **c**, Density of NB001 in faeces before and after addition of inulin in the diet (orange shading) RF, $n = 4$; Hum-1, $n = 4$; Hum-2, $n = 4$; Kruskal–Wallis test, $^{*}P = 0.03$. **d**, Density of NB001 in faeces before and after addition of seaweed in the diet (green shading). RF, $n = 5$; Hum-1, $n = 7$; Hum-2, $n = 6$; Kruskal–Wallis test, $P = 0.19$ (NS, not significant). Data are mean ± s.e.m. The grey-shaded boxes represent the limit of detection.

---

[1]Department of Microbiology and Immunology, Stanford University School of Medicine, Stanford, CA, USA. [2]Novome Biotechnologies, South San Francisco, CA, USA. *e-mail: jsonnenburg@stanford.edu

ex-germ-free mice, each colonized with the gut microbiota from a different healthy human donor from the United States (humanized), as model hosts. These three groups of mice received a rare strain of the prominent gut commensal species *Bacteroides ovatus* (NB001), which we isolated specifically for its ability to utilize both dietary fructans and marine polysaccharides (discussed below). NB001 was monitored by green-fluorescent protein (GFP)-positive[12] colony forming units (c.f.u.) in faeces via selective plating over the course of seven days (Fig. 1a). The different communities (Extended Data Fig. 1) varied in their ability to integrate NB001 (range of mean log$_{10}$(c.f.u.) per ml in faeces on day 7 of < 4.60–6.70; Fig. 1b) and one human microbiota was resistant to colonization altogether.

Diet is a primary selective force that shapes community membership and functionality[13–16]. Members of the commensal genus *Bacteroides* are prolific utilizers of myriad diet-derived microbiota-accessible carbohydrates (MACs) through the machinery encoded by their polysaccharide utilization loci (PULs)[17,18]. Given the competition for resources in the gut, we hypothesized that specific MACs may serve as a route through which colonization can be modulated across varied communities. Previously, we used this strategy in gnotobiotic mice colonized with two species of *Bacteroides*, using the dietary MAC inulin to enable the inulin-using strain to proliferate[19]. To test the applicability of this approach in the context of a complex microbiota, we administered inulin to the three aforementioned groups of mice seven days after inoculation with NB001, which is capable of robust growth on inulin as the only carbon source (Extended Data Fig. 2a). Over seven days of feeding the same inulin-based diet used in our previous study (10% inulin w/w) to these three groups of mice, NB001 exhibited variable responses across the background microbiotas (range of mean log$_{10}$(c.f.u.) per ml in faeces 7 days after diet change of < 4.60–6.67; Fig. 1c).

By contrast, microbiota utilization of the non-ubiquitous marine polysaccharide porphyran, the primary carbohydrate in the seaweed *Porphyra yezoensis* used to prepare culinary nori, is much less common in US microbiotas[15]. The consumption of porphyran by rare *Bacteroides* strains is enabled by a horizontally transferred PUL, which originated in marine bacteria[15,20]. Whole-genome sequencing of NB001 revealed a gene cluster highly homologous to a previously described porphyran PUL (Extended Data Fig. 2b). We verified the requirement of the intact PUL for specific growth on porphyran in vitro by knockout of porphyran utilization genes (Extended Data Fig. 2c). We hypothesized that this unique dietary MAC would create a privileged niche within the gut and promote engraftment of an exogenous strain competent in its use. Therefore, we administered a custom diet supplemented with porphyran-rich seaweed (10% nori w/w) to the three groups of mice seven days after inoculation with NB001. Indeed, in response to seaweed in the diet, we observed a robust increase in abundance of NB001 (four to six orders of magnitude), irrespective of background microbiota. Additionally, the variability in colonization levels across communities was eliminated (range of mean log$_{10}$(c.f.u.) per ml in faeces 7 days after diet change of 10.34–10.96; Fig. 1d), indicating specificity of a privileged nutrient source and its cognate utilization system to promote bacterial growth in vivo. Access to porphyran in the diet rescued NB001 from below the limits of detection in mice with the most resistant microbiota and boosted its abundance to levels indistinguishable from those achieved in mice with the other two microbiotas (Fig. 1d). Despite potential limitations of this specific polysaccharide system in populations colonized with competing porphyran utilizers (for example, a small subset of Japanese individuals[15]), together, these data suggest a powerful approach in using rare pairs of nutrients and their cognate utilization systems to reliably control strain engraftment independent of the background microbiota, and implicate nutrient availability as a key modulator of strain integration into the gut community.

Given the context independence of strain engraftment via access to porphyran in the diet (Fig. 1d), we next tested whether the population size could be reversibly expanded by addition and removal of the substrate in vivo. We colonized mice that had conventional microbiota with NB001 and tracked the c.f.u. in faeces before, during and after a five-day administration of 1% porphyran in the drinking water. To test the influence of competing polysaccharides in the diet, we performed this experiment on both standard, MAC-rich chow and a MAC-deficient chow that provides no exogenous polysaccharides to the microbiota. With or without other competing dietary MACs, NB001 responded robustly to introduction of porphyran, showing a large and highly reproducible increase in abundance (Fig. 2a, b). This response was contingent upon access to porphyran, as deletion of eight genes required for its metabolism abolished the effect (Fig. 2c). Additionally, the porphyran alone did not significantly affect the composition of the background microbiota (Extended Data Fig. 3), supporting the lack of porphyran use by members of the background community.

Bacteroides species engage in interesting colonization behaviour in which an early colonizer will exclude a challenging isogenic strain[4], a phenomenon known as priority effects in the ecological literature[21,22]. We hypothesized that this behaviour could be overcome by providing a privileged nutrient to the challenging strain to create an exclusive

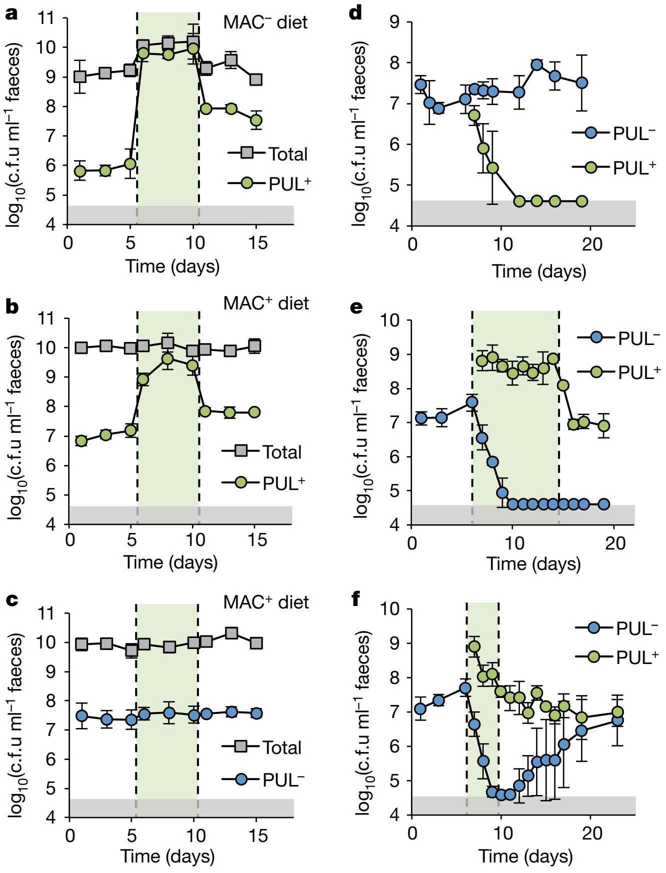


Fig. 2 | Access to a privileged nutrient mediates population size and overcomes isogenic self-exclusion. Density of NB001 (PUL^{+}), NB001 lacking the ability to utilize porphyran (PUL^{-}) and total culturable anaerobes (Total) in faeces of mice with conventional microbiota. Periods of administration of porphyran in drinking water are indicated by green shading (**a**–**c**, 1% w/v (maximum possible); **d**–**f**, 0.1% w/v (to induce 1-log above PUL^{-}). **a**, **b**, Mice colonized with PUL^{+} and fed a MAC-deficient (MAC^{-}) diet (**a**) or MAC-rich (MAC^{+}) diet (**b**) demonstrated reversible expansion of PUL^{+} upon administration of porphyran. **c**, Mice colonized with PUL^{-} and fed a MAC-rich diet showed no change in strain abundance with administration of porphyran. **d**–**f**, Mice that were fed a MAC-rich diet were colonized with PUL^{-} for six days, and challenged with PUL^{+} on day 6. **d**, PUL^{+} is excluded by PUL^{-} in the absence of porphyran. **e**, PUL^{+} displaces PUL^{-} with access to porphyran for eight days. **f**, PUL^{+} and PUL^{-} stably co-exist after a three-day pulse of porphyran. Data are mean ± s.d.; $n = 4$ for all experiments. The grey-shaded boxes represent the limit of detection.

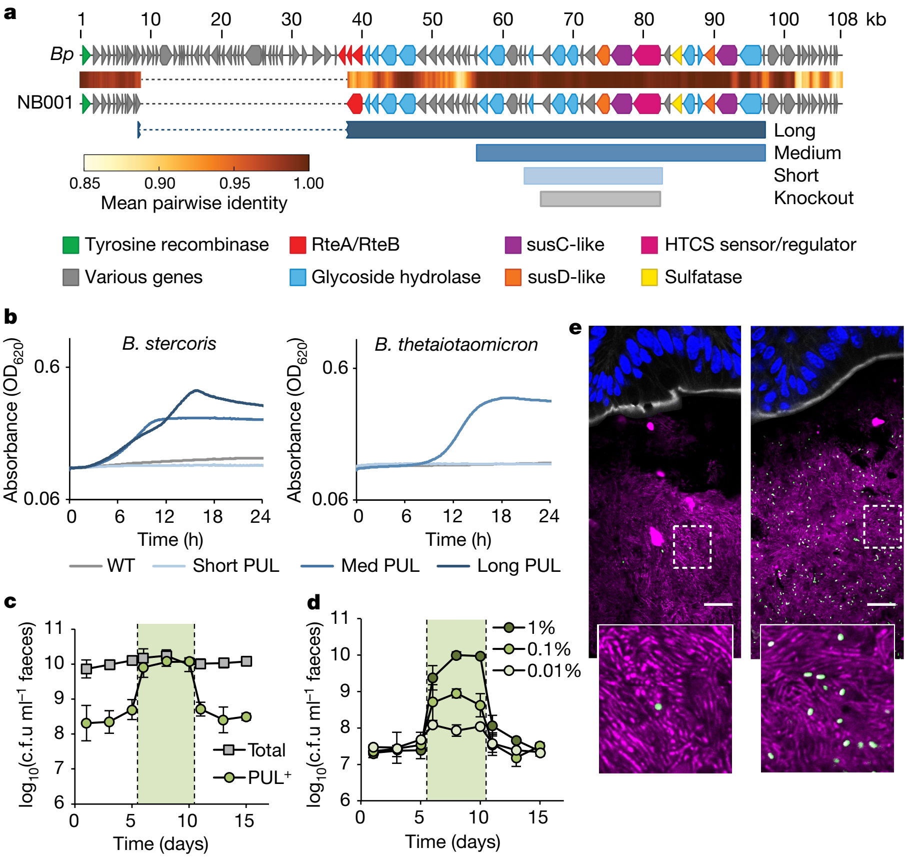


Fig. 3 | Control over population size can be engineered and is highly tunable. a, Schematic of porphyran PUL from NB001 aligned to the previously published *B. plebeius* (*Bp*) PUL. The different minimal PULs (long = 34 genes, medium = 21 genes, and short = 10 genes) designed and tested for ability to confer growth on porphyran are shown as blue bars. The eight-gene region deleted from NB001 PUL$^-$ is shown in grey (knockout). **b**, Growth curves for wild-type and engineered strains on porphyran as the only carbon source. **c**, *B. thetaiotaomicron* with the medium length PUL colonized in mice with conventional microbiota fed a MAC-rich diet demonstrates reversible expansion upon administration of 1% w/v porphyran in the drinking water (shaded green, *n* = 4). Data are mean ± s.d. **d**, NB001 PUL$^+$ colonized in mice with conventional microbiota consuming a MAC-rich diet demonstrates tunable (finely controlled) response to porphyran in the drinking water (shaded green, *n* = 4 per group). Data are mean ± s.d. **e**, NB001 expressing GFP colonized in conventional mice that received 0.01% porphyran (left) or 1% porphyran (right). Image is of proximal colon with host epithelium visualized by DAPI (nuclei, blue), epithelial border visualized by phalloidin (F-actin, white), background microbiota by DAPI segmented from host epithelium (bacteria, magenta), NB001 by endogenous GFP (bacteria, green). Scale bars, 16 μm.

metabolic niche. Indeed, an early-colonizing version of NB001 lacking the functional porphyran PUL excludes a PUL-competent NB001 challenge strain in mice with conventional microbiota (Fig. 2d). Supplementing porphyran in the water, accessible only to the challenging strain, enabled the challenging strain to overcome the priority effect, and resulted in displacement of the early colonizer by the challenging strain (Fig. 2e). Replacement of the early-colonizing strain required that the challenging strain had access to porphyran (Extended Data Fig. 4a), and displacement was robust to subsequent challenge by the original colonizer (Extended Data Fig. 4b). Interestingly, an intermediate colonization state could be achieved with shorter porphyran administration (Fig. 2f).

On the basis of the sequence of the NB001 porphyran PUL, using alignment to the previously published PUL from *Bacteroides plebeius*[15,20] (Fig. 3a), we constructed three different minimal PULs, ranging from 10 to 34 of the genes (20–60 kb) from the full-length PUL (Fig. 3a). Previous methods[19] used to transfer a five-gene PUL were not sufficient for these much larger constructs, and so we assembled the plasmids via homologous recombination in yeast[23] before integration into the chromosome of two target strains unable to utilize porphyran for growth, *Bacteroides stercoris* and *Bacteroides thetaiotaomicron* (Extended Data Fig. 5). Transfer of the short 10-gene PUL was insufficient to impart growth on porphyran to either of the naive species of *Bacteroides* tested (Fig. 3b), but the medium (21-gene) and long (34-gene) PULs both enabled growth of strains in vitro (Fig. 3b). Notably, the two species harbouring the medium 21-gene PUL reached different maximum absorbance during growth, although this could be due to differences in the efficiency of growth in minimal medium (Extended Data Fig. 6). A substantial drop-off in conjugation efficiency was observed with the 34-gene PUL, and only *B. stercoris* yielded transconjugants. When colonized into mice with conventional microbiota, abundance of *B. thetaiotaomicron* harbouring the medium 21-gene PUL could be reversibly expanded through addition of porphyran in the water (Fig. 3c), and this also occurred when mice were colonized with *B. stercoris* harbouring the long 34-gene PUL (Extended Data Fig. 7).

We next sought to determine whether strain abundance could be tuned by varying the amount of porphyran supplemented in the diet. We colonized mice that had conventional microbiota with a GFP fluorescent NB001 strain and tracked c.f.u. in the faeces before, during and after administration of 1%, 0.1% or 0.01% w/v porphyran in the drinking water (resulting in an estimated dose of 70, 7, or 0.7 mg porphyran per mouse per day, respectively). For each tenfold dilution of porphyran administered in the water, we observed a tenfold decrease in abundance of NB001 in the faeces (Fig. 3d), indicating fine control over strain abundance through modulation of the porphyran concentration. When examining prepared frozen tissue sections of proximal colon from mice consuming 0.01% or 1% porphyran in the water for endogenous GFP of the NB001 strain via confocal microscopy, a substantial increase in GFP-positive cells

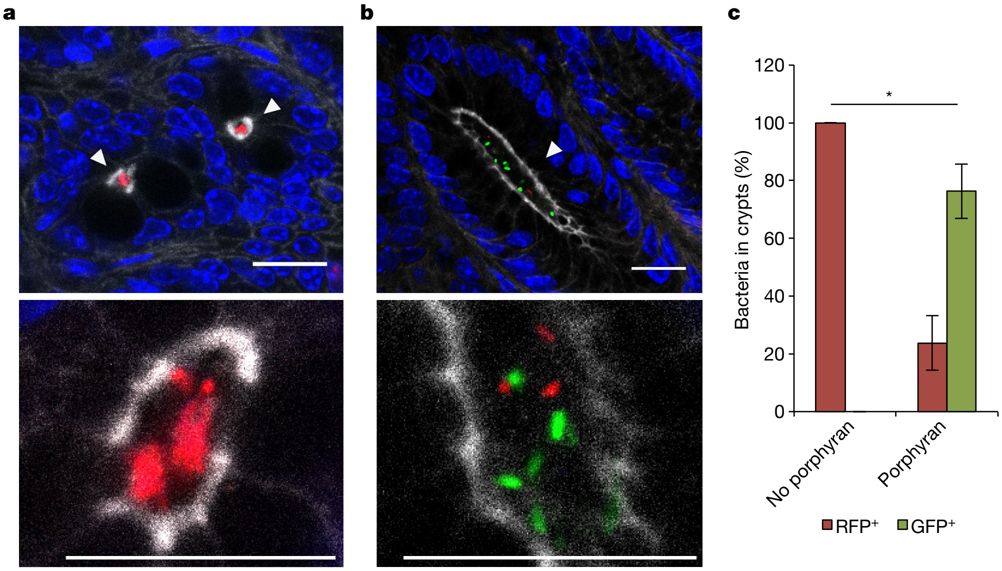


Fig. 4 | Access to porphyran enables colonization of the colonic crypt niche. Wild-type *B. thetaiotaomicron* expressing RFP was introduced into germ-free mice seven days before challenge with engineered *B. thetaiotaomicron* harbouring the 21-gene PUL and expressing GFP. The challenge strain of *B. thetaiotaomicron* is excluded from colonic crypts when no porphyran is administered (**a**), but can remodel the crypt population when given access to porphyran through seaweed in the diet (**b**). Images are proximal colon with host epithelium visualized by DAPI (nuclei, blue), epithelial border including crypt borders visualized by phalloidin (F-actin, white), wild-type *B. thetaiotaomicron* by endogenous RFP (bacteria, red) and PUL$^+$ *B. thetaiotaomicron* by endogenous GFP (bacteria, green). Bacteria in crypts are highlighted by white arrowheads. Scale bars, 16 μm. **c**, Quantification of bacteria in crypts in the presence or absence of porphyran through seaweed in the diet. $n = 3$ mice per group, $n > 100$ crypt-resident bacteria counted per mouse; two-tailed t-test, $^*P = 0.0395$. Data are mean ± s.d.

was seen when comparing mice that received 0.01% compared to 1% porphyran (Fig. 3e).

Finally, we investigated whether access to porphyran enabled colonization of the colonic crypts. We previously observed that *B. thetaiotaomicron* is excluded from crypt colonization when an isogenic strain of *B. thetaiotaomicron* is already established in the gut[12], and sought to overcome this crypt exclusion with privileged access to porphyran. We colonized germ-free mice with wild-type *B. thetaiotaomicron* expressing red-fluorescent protein (RFP) and challenged seven days later with an engineered *B. thetaiotaomicron* harbouring the 21-gene porphyran PUL and expressing GFP. When mice received no porphyran in the diet, the GFP-expressing strain was excluded as anticipated (Fig. 4a, c). However, when mice received porphyran-rich seaweed chow, the GFP-expressing, porphyran-using strain was able to colonize the colonic crypts (Fig. 4b, c), suggesting privileged nutrient-utilization enables remodelling of this specialized microhabitat.

Factors governing commensal niche availability in the gut microbiota are not well defined, and predicting the amenability of a given community to the introduction of a new strain is a challenge. Here we have illustrated the concept of exclusive nutrient access as an important principle underlying strain engraftment and abundance control. Moreover, the context independence, reversibility and tunability of this approach expand the utility of these results to broad applications ranging from experimental manipulation of complex communities to fine-tuned control of therapeutic strains or cocktails of microorganisms. Understanding the role of nutrient utilization in the microbiota clarifies our image of the gut niche landscape, and will help to facilitate the design of systems that can be deployed to enhance human health.

Online content

Any Methods, including any statements of data availability and Nature Research reporting summaries, along with any additional references and Source Data files, are available in the online version of the paper at https://doi.org/10.1038/s41586-018-0092-4.

Received: 22 July 2017; Accepted: 23 March 2018;
Published online 9 May 2018.

1. Faith, J. J. et al. The long-term stability of the human gut microbiota. *Science* **341**, 1237439 (2013).
2. Frese, S. A., Hutkins, R. W. & Walter, J. Comparison of the colonization ability of autochthonous and allochthonous strains of lactobacilli in the human gastrointestinal tract. *Adv. Microbiol.* **2**, 399–409 (2012).
3. Maldonado-Gómez, M. X. et al. Stable engraftment of *Bifidobacterium longum* AH1206 in the human gut depends on individualized features of the resident microbiome. *Cell Host Microbe* **20**, 515–526 (2016).
4. Lee, S. M. et al. Bacterial colonization factors control specificity and stability of the gut microbiota. *Nature* **501**, 426–429 (2013).
5. Ivanov, I. I. et al. Induction of intestinal Th17 cells by segmented filamentous bacteria. *Cell* **139**, 485–498 (2009).
6. Atarashi, K. et al. Induction of colonic regulatory T cells by indigenous *Clostridium* species. *Science* **331**, 337–341 (2011).
7. Lawley, T. D. et al. Targeted restoration of the intestinal microbiota with a simple, defined bacteriotherapy resolves relapsing *Clostridium difficile* disease in mice. *PLoS Pathog.* **8**, e1002995 (2012).
8. Stecher, B. et al. Like will to like: abundances of closely related species can predict susceptibility to intestinal colonization by pathogenic and commensal bacteria. *PLoS Pathog.* **6**, e1000711 (2010).
9. Ratner, M. Seres's pioneering microbiome drug fails mid-stage trial. *Nat. Biotechnol.* **34**, 1004–1005 (2016).
10. Ericsson, A. C., Personett, A. R., Turner, G., Dorfmeyer, R. A. & Franklin, C. L. Variable colonization after reciprocal fecal microbiota transfer between mice with low and high richness microbiota. *Front. Microbiol.* **8**, 196 (2017).
11. Landy, J. et al. Variable alterations of the microbiota, without metabolic or immunological change, following faecal microbiota transplantation in patients with chronic pouchitis. *Sci. Rep.* **5**, 12955 (2015).
12. Whitaker, W. R., Shepherd, E. S. & Sonnenburg, J. L. Tunable expression tools enable single-cell strain distinction in the gut microbiome. *Cell* **169**, 538–546 (2017).
13. Sonnenburg, E. D. et al. Diet-induced extinctions in the gut microbiota compound over generations. *Nature* **529**, 212–215 (2016).
14. Ridaura, V. K. et al. Gut microbiota from twins discordant for obesity modulate metabolism in mice. *Science* **341**, 1241214 (2013).
15. Hehemann, J. H. et al. Transfer of carbohydrate-active enzymes from marine bacteria to Japanese gut microbiota. *Nature* **464**, 908–912 (2010).
16. David, L. A. et al. Diet rapidly and reproducibly alters the human gut microbiome. *Nature* **505**, 559–563 (2014).
17. Sonnenburg, J. L. et al. Glycan foraging in vivo by an intestine-adapted bacterial symbiont. *Science* **307**, 1955–1959 (2005).
18. Xu, J. et al. Evolution of symbiotic bacteria in the distal human intestine. *PLoS Biol.* **5**, e156 (2007).
19. Sonnenburg, E. D. et al. Specificity of polysaccharide use in intestinal *Bacteroides* species determines diet-induced microbiota alterations. *Cell* **141**, 1241–1252 (2010).
20. Hehemann, J. H., Kelly, A. G., Pudlo, N. A., Martens, E. C. & Boraston, A. B. Bacteria of the human gut microbiome catabolize red seaweed glycans with carbohydrate-active enzyme updates from extrinsic microbes. *Proc. Natl Acad. Sci. USA* **109**, 19786–19791 (2012).

21. Drake, J. A. Community-assembly mechanics and the structure of an experimental species ensemble. *Am. Nat.* **137**, 1–26 (1991).
22. Fukami, T. historical contingency in community assembly: integrating niches, species pools, and priority effects. *Annu. Rev. Ecol. Evol. Syst.* **46**, 1–23 (2015).
23. Chandran, S. & Shapland, E. in *Synthetic DNA* Vol. 1472 (ed. Hughes, R. A.) 187–192 (Humana, New York, 2017).

Acknowledgements We thank N. Ratnayeke for early experimental assistance, S. Higginbottom for gnotobiotic assistance, and K. Ng, Z. Russ and W. Van Treuren for analytical assistance; the Amieva and Huang laboratories for use of their microscopy resources, and D. Shepherd for valuable discussions; N. Pudlo and E. Martens for the protocol on porphyran extraction, and E. Sonnenburg, T. Fukami and M. Fischbach for commenting on this manuscript. This material is based upon work supported by the National Science Foundation under grant number 1648230, the NIDDK (R01-DK085025 to J.L.S.) and an NSF Graduate Fellowship (DGE-114747 to E.S.S.).

Reviewer information *Nature* thanks D. Bolam and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions E.S.S. and J.L.S. generated the idea for the project. E.S.S. performed the in vivo studies, 16S sample preparation and analysis and in vitro growth studies. W.C.D. sequenced and analysed the NB001 genome and constructed the porphyran PUL plasmids. K.M.P. and E.S.S. performed in vivo crypt studies, imaging and analysis. W.R.W. isolated NB001 and performed in vitro growth characterization. All authors contributed to experimental design and data analysis, and E.S.S. and J.L.S. wrote the manuscript. All authors discussed the results and commented on the manuscript.

Competing interests E.S.S., W.C.D., W.R.W. and J.L.S. are founders at Novome Biotechnologies, Inc. and have filed a provisional patent based on the work described here (US Provisional Patent No. 62/435,048).

Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0092-4.
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0092-4.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to J.L.S.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

Bacterial culture and strain isolation. All bacterial growth was performed at 37 °C under anaerobic conditions. Growth for introduction into mice was performed in rich medium (tryptone–yeast–glucose[12]) with no antibiotics added. Growth curves and selective growth on porphyran were performed in Salyers minimal medium (SMM 100 ml in dH_2O: 0.1 g $(NH_4)_2SO_4$, 0.1 g Na_2CO_3, 0.05 g L-cysteine, 10 ml 1 M KPO_4 pH 7.2, 5 ml mineral salts (1 l in dH_2O: 18 g NaCl, 0.53 g $CaCl_2{\cdot}2H_2O$, 0.4 g $MgCl_2{\cdot}6H_2O$, 0.2 g $MnCl_2{\cdot}4H_2O$, 0.2 g $CoCl_2{\cdot}6H_2O$), 1 ml 0.4 mg ml^{-1} $FeSO_4$, 0.1 ml 1 mg ml^{-1} vitamin K_3, 0.1 ml histidine/haematin (12 mg haematin in 10 ml of 0.2 M histidine, pH 8), 0.05 ml 0.01 mg ml^{-1} vitamin B_{12}) with a carbon source added to a final concentration of 0.2% (Extended Data Fig. 5), 0.5% (Extended Data Figs. 2) or 0.8% (Fig. 3b) w/v and filter-sterilized. Path length for our growth curves was 0.58 cm. SMM was either made fresh on the day of the experiment, or prepared without histidine/haematin or L-cysteine and stored at 4 °C for up to four weeks, with the missing components added on the day of the experiment.

The c.f.u. was determined by serial dilution of faeces (resuspension of 1 μl of fresh faeces in PBS, performing serial dilutions, and extrapolating to c.f.u. ml^{-1} faeces) and culturing on brain–heart infusion blood agar (BHI-BA) with appropriate selective antibiotics (200 μg ml^{-1} gentamicin, and 25 μg ml^{-1} erythromycin or 2 μg ml^{-1} tetracycline), or on SMM agar plates (for GFP visualization, 2× SMM was heated to 50 °C and combined with an equal volume 3% agar heated to 50 °C) with selective antibiotics. Total culturable anaerobe number was determined via plating serial dilutions of faecal matter on BHI-BA without antibiotics.

NB001 (porphyran using *B. ovatus*) and NB004 (naive *B. stercoris*) were isolated from primary waste effluent at the San Jose Wastewater Treatment Facility via selection in liquid culture (SMM) with 200 μg ml^{-1} gentamicin (as *Bacteroides* are naturally gentamicin-resistant), and for NB001, growth on 0.8% porphyran. In brief, settled primary effluent was diluted tenfold into SMM and grown as above for 24 h, subcultured at 1:200 into fresh medium and grown for 24 h, and plated in serial dilutions onto BHI-BA. Single colonies were picked into SMM for growth confirmation, cryogenic storage and downstream analysis. Species were identified via whole-genome sequencing. A GFP-expressing, erythromycin-resistant variant of NB001 was generated as described previously[12].

Genome sequencing and analysis. Genomic DNA was isolated from NB001 and NB004 using a PureLink Genomic DNA Mini Kit (Invitrogen). Samples were prepared for multiplexed Illumina sequencing using a Nextera XT DNA Library Preparation Kit (Illumina) and run on an Illumina MiSeq using a 2 × 150-bp paired-end kit. Approximately 10 million sequencing reads were obtained for each sample. De novo assembly of the reads was performed with the Geneious De novo Assembler (Biomatters), yielding an average coverage of around 100 reads per bp. Gene annotation and alignment was also performed using Geneious (https://www.geneious.com/).

Porphyran PUL transfer and knockout. Generation of the porphyran utilization deficient mutant was performed using *tdk* counterselection as described previously[24]. A thymidine kinase (*tdk*) deficient mutant (NB007) of NB001 was generated by exposing an aliquot of liquid culture of NB001 to 320-nm ultraviolet light from a VWR-20E transilluminator (VWR) for 60 s and plating on BHI-BA supplemented with 200 μg ml^{-1} of 5-fluoro-2′-deoxyuridine (FUdR). Eight genes predicted to be essential for growth on seaweed MACs (homologous to BACPLE_1692-1699) were knocked out using pWD034 (Extended Data Table 1), a plasmid with 1.5 kb of homology upstream and downstream of the target region was assembled via Golden Gate Assembly into an erythromycin-resistant, *tdk*-containing vector.

Generation of the seaweed PUL knock-in strains required expansion of previous knock-in methods[19] because of the large size of the PULs (20—60 kb). On the basis of gene annotations and sequence alignment to a previously published mobile element conferring seaweed polysaccharide utilization capabilities[15,20], we designed three minimal PULs of varying sizes (20 kb, 40 kb and 60 kb; Fig. 3a). To propagate such large pieces of DNA and integrate them into the *Bacteroides* genome, we used a three-step process: performing yeast assembly into a custom shuttle vector, propagating it in *Saccharomyces cerevisiae* and *Escherichia coli*, and then performing conjugation and genomic integration into *Bacteroides* species. The minimal PULs were each divided into multiple 6-kb fragments with 200-bp homology between pieces and assembled in yeast[23] with fragments containing the KanMX selectable marker and CEN6/ARS4 origin for selection and growth in yeast, the bacterial artificial chromosome origin and chloramphenicol selectable marker (from the pEZ-BAC vector, Lucigen) for selection and growth in *E. coli*, and the conjugative origin and parts for integration and selection in *Bacteroides* (Supplementary Information). Yeast cells with successfully assembled constructs were lysed by mechanical disruption with 0.5-mm glass disruptor beads (USA Scientific), and lysates (raw lysates not purified for plasmid DNA) were mixed 1:20 with electrocompetent *E. coli* S17-1 cells[25]. DNA was introduced to the *E. coli* cells via electroporation according to a previously described protocol[26]. Successfully transformed *E. coli* were then grown and conjugated with NB004 and *B. thetaiotaomicron* VPI-5482 as previously described[12]. NB004 successfully integrated all PULs, but *B. thetaiotaomicron* VPI-5482 was initially unable to integrate any PUL constructs. To improve rates of genomic integration, we pre-integrated an NBU2 integrase-expressing plasmid with tetracycline resistance into *B. thetaiotaomicron* VPI-5482, and this improved efficiency such that *B. thetaiotaomicron* VPI-5482 conjugants were obtained for both the short and medium length PUL constructs.

Porphyran preparation. Raw culinary nori derived from *Porphyra yezoensis* (acquired from https://www.rawnori.com) was added at 10% w/v to distilled water and subjected to hot water extraction by autoclaving for three hours. The mixture was then cooled and centrifuged at 11,000 *g*. The resulting supernatant was ethanol precipitated by combining the solution with 100% ethanol to a final concentration of 80% ethanol, 20% supernatant and incubating at 4 °C for 24–72 h. The precipitate was recovered by centrifugation at 26,000 *g* and dried for 24 h before manual grinding to generate a measurable powder.

Mice. Ex-germ-free or conventionally colonized (conventional, restricted flora (RF)) male or female Swiss-Webster mice (Taconic) aged 8–16 weeks were housed in gnotobiotic isolators and fed either an autoclaved standard diet (LabDiet 5K67) or a custom diet as indicated below (Bio-Serv) in strict accordance with a Protocol for Care and Use of Laboratory Animals approved by the Stanford University Administrative Panel of Laboratory Animal Care. Sample size was chosen to generate enough power for statistical significance, animals were split into groups at time of weaning, blinding was not performed and the experiments were not randomized. Introduction of all *Bacteroides* strains into either ex-germ free or RF mice was performed by oral gavage of 10^8 c.f.u. of the given strain in culture medium.

Mice were humanized (Fig. 1) with faecal samples from healthy human donors that were stored at −80 °C, thawed and resuspended in pre-reduced PBS in anaerobic conditions at a 1:1 dilution, and 0.2 ml administered orally into germ-free mice (total sample administered per mouse equivalent to 0.1 ml frozen faecal matter). The mice were allowed to equilibrate the human microbiota for four weeks while consuming the standard laboratory diet (Purina LabDiet 5K67, mix of crude protein, fat and carbohydrates). One week before introduction of NB001, mice were switched to MAC-deficient chow (AIN-93G, 68% glucose). Seven days after introduction of NB001, mice were switched to custom diets with inulin or seaweed as the only available MACs (AIN-93G, 10% unique polysaccharide, 58% glucose). RF mice (Figs. 2, 3) were administered porphyran in the water at the percentage indicated (w/v). Gnotobiotic mice were fed either standard laboratory diet (Fig. 4a) or custom seaweed chow (Fig. 4b) as above.

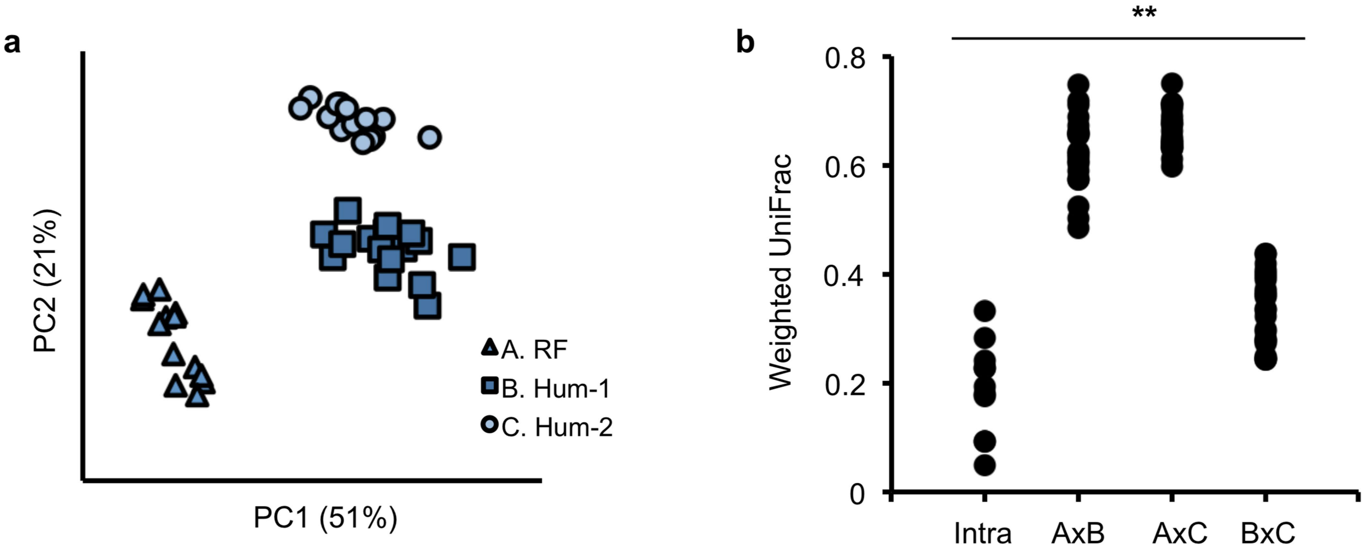
16 S rRNA analysis. DNA was extracted from faecal samples using the PowerSoil 96-htp kit (MoBio), and amplified at the 16 S v4 region (515 F, 806 R). Qiime 1.9 was used to analyse the resulting Illumina-generated sequencing reads as previously described[27]. Data were rarified to the sample with the lowest number of reads (16,384), and open-reference OTU picking via UCLUST and taxonomy assignment through the Greengenes 13.8 database (http://qiime.org/home_static/dataFiles.html) was performed.

Microscopy. Tissue was harvested and immediately fixed in 4% paraformaldehyde in PBS for 48 h at 4 °C. Cassettes were transferred to 20% sucrose for 24 h, and then samples were embedded in OCT compound (Tissue-Tek) before sectioning at 30 μm on a Leica CM3050S cryostat. Sections were stained for 30 min with 4′,6-diamidino-2-phenylindole dihydrochloride (DAPI; Sigma-Aldrich) and Alexa Fluor 594 phalloidin (Life Technologies). Images were taken on a Zeiss LSM 700 confocal microscope.

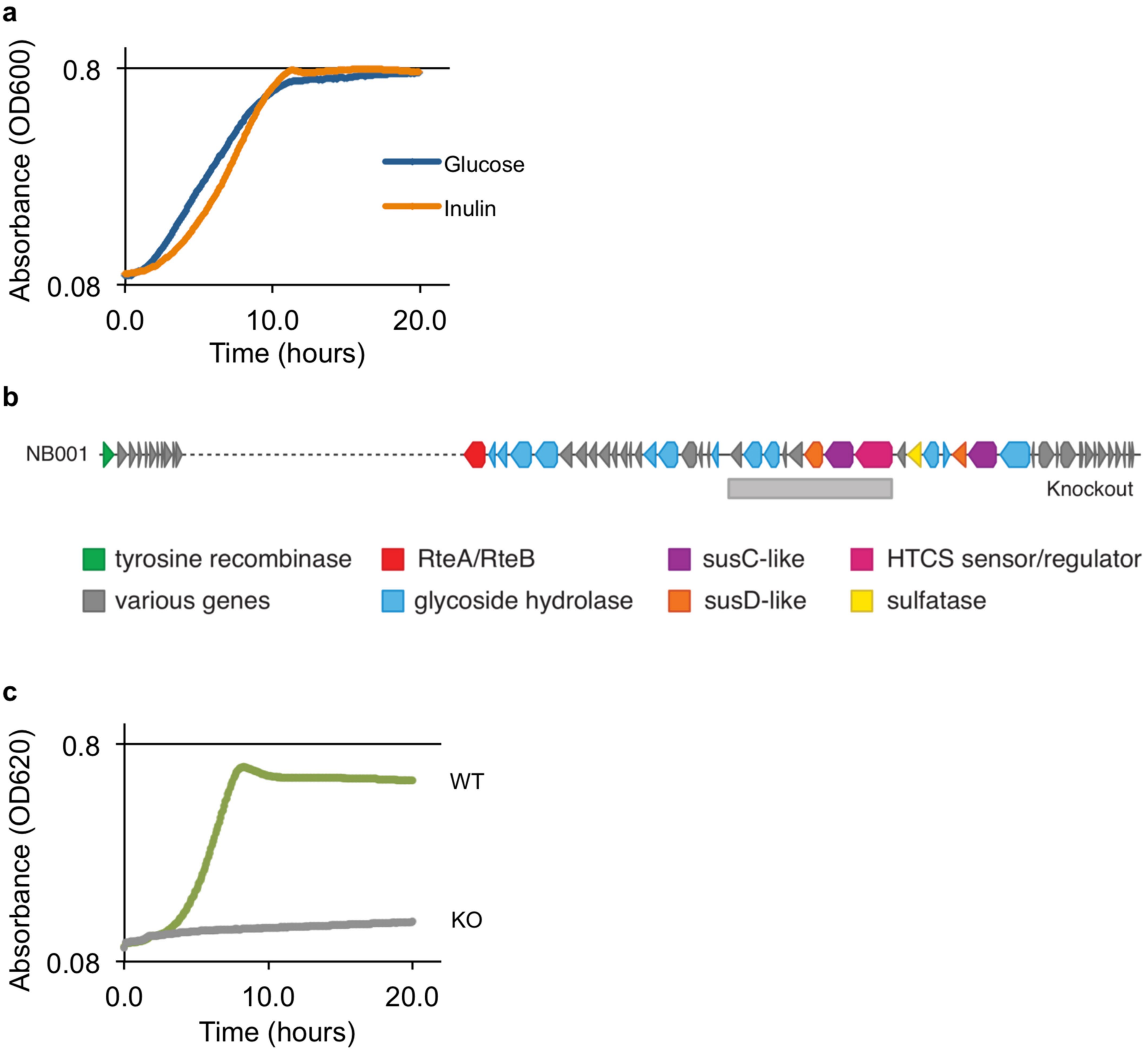
Reporting summary. Further information on experimental design is available in the Nature Research Reporting Summary linked to this paper.

Data availability. The 16 S sequencing data have been deposited in the Sequence Read Archive under accession number PRJNA436622. Source Data for all animal experiments are available in the online version of this paper.

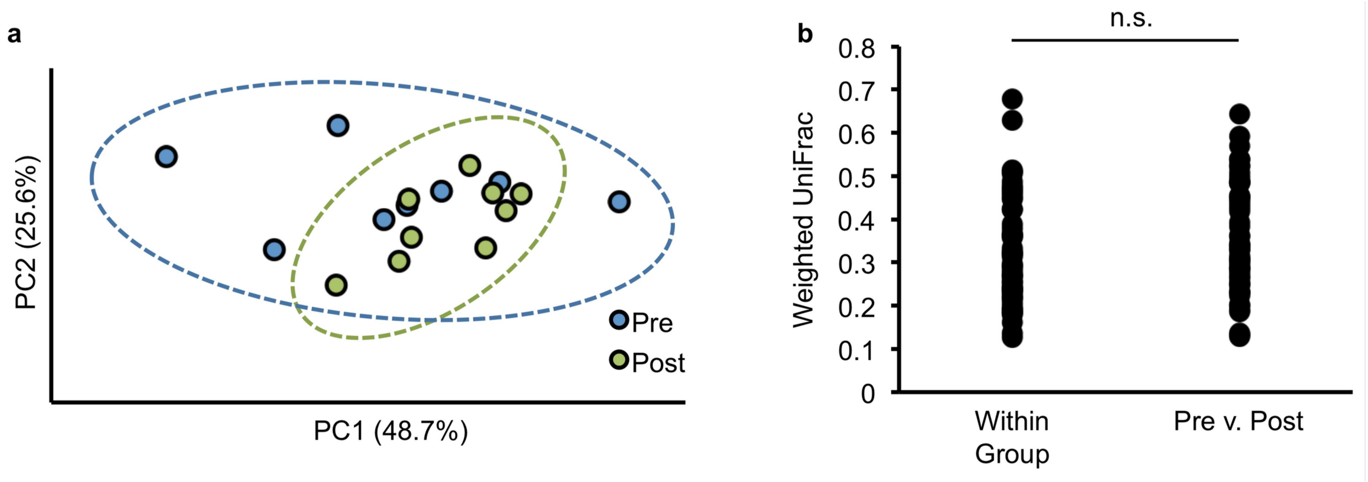
24. Martens, E. C., Chiang, H. C. & Gordon, J. I. Mucosal glycan foraging enhances fitness and transmission of a saccharolytic human gut bacterial symbiont. *Cell Host Microbe* **4**, 447–457 (2008).
25. Simon, R., Priefer, U. & Pühler, A. A broad host range mobilization system for in vivo genetic engineering: transposon mutagenesis in gram negative bacteria. *Nat. Biotechnol.* **1**, 784–791 (1983).
26. Ostrov, N. et al. Design, synthesis, and testing toward a 57-codon genome. *Science* **353**, 819–822 (2016).
27. Caporaso, J. G. et al. QIIME allows analysis of high-throughput community sequencing data. *Nat. Methods* **7**, 335–336 (2010).



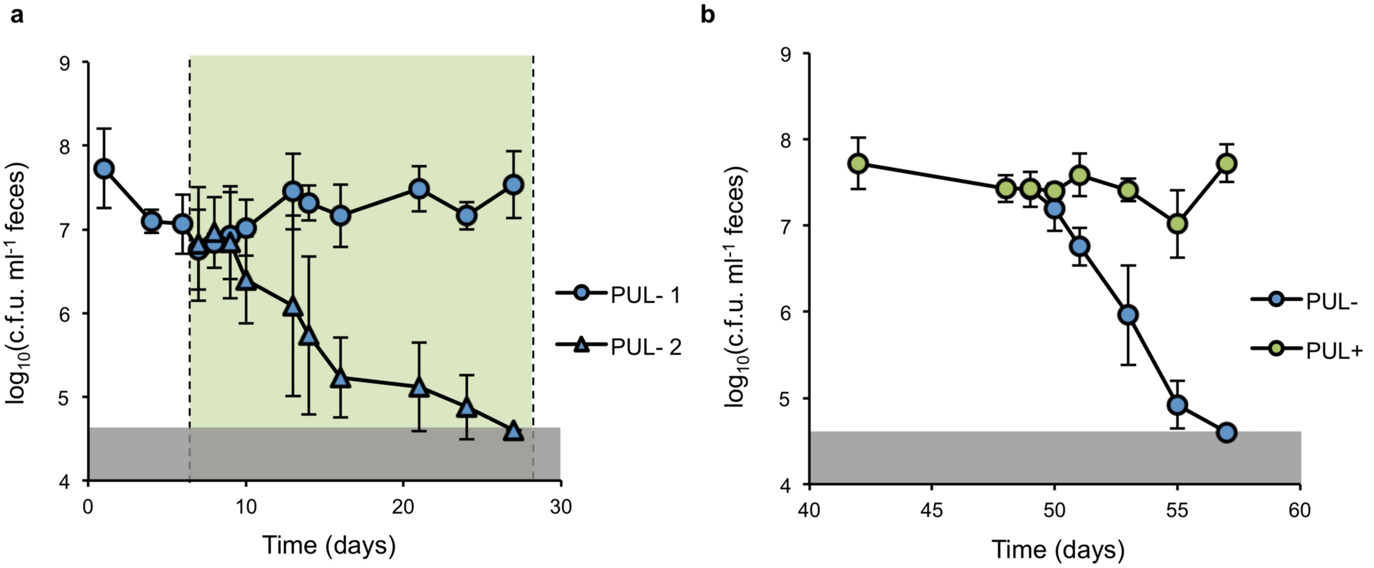
Extended Data Fig. 1 | Three model background communities of gut microbes are distinct from each other. a, Principal coordinates (PC) analysis of weighted UniFrac distance for 16 S rRNA gene amplicons from faeces from the three background community groups from Fig. 1 before diet switch for conventional (RF) or humanized (Hum-1 and Hum-2) mice, *n* = 10. **b**, Comparison of weighted UniFrac distances within each group (Intra) or across groups (A × B, A × C, B × C). One-way ANOVA, *P* < 0.0001.



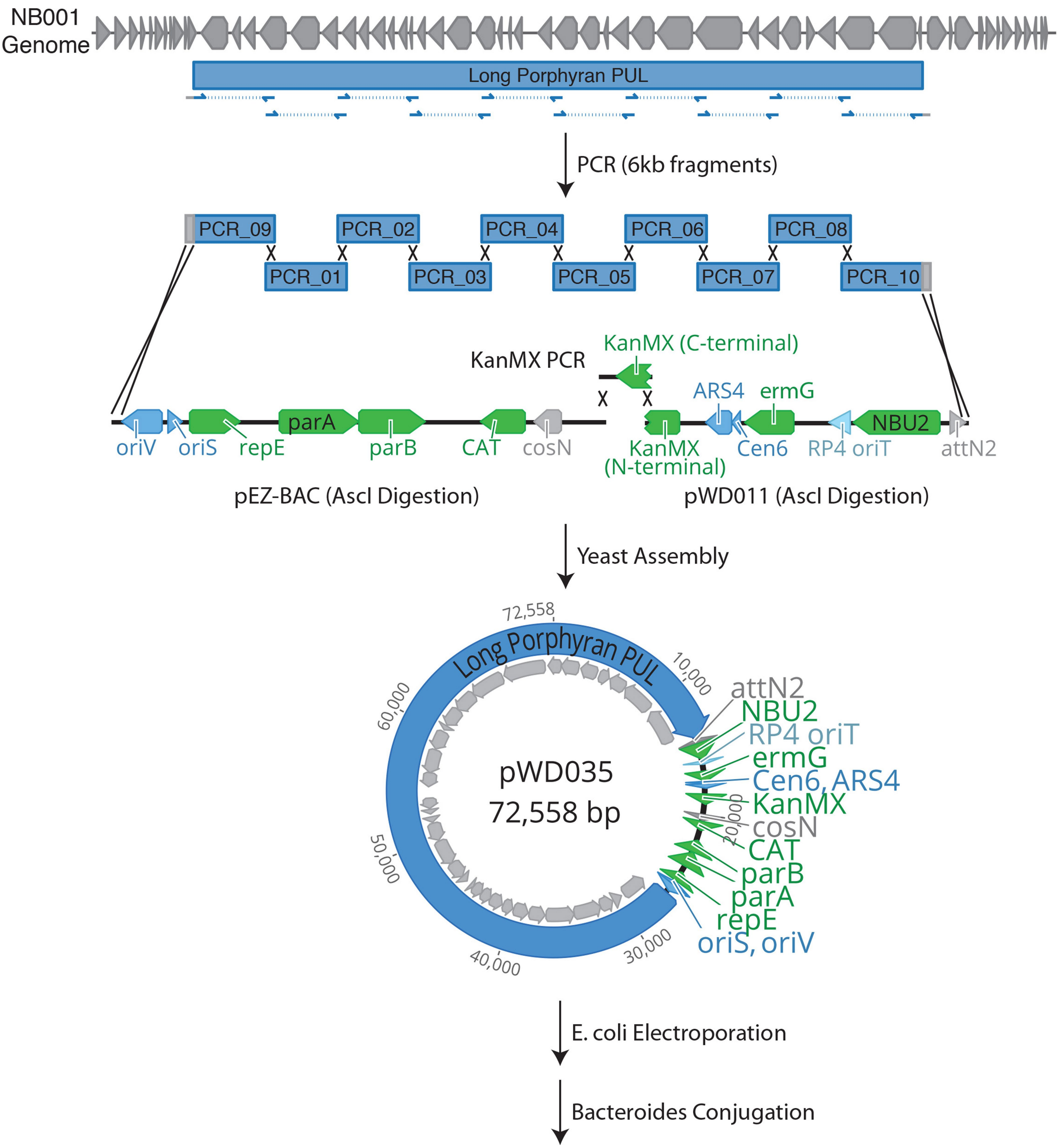
Extended Data Fig. 2 | NB001 can utilize both inulin and porphyran as the only carbon source for growth. a, NB001 demonstrates growth in minimal medium with either glucose (blue, doubling time = 157 min), or inulin (orange, doubling time = 127 min), as the only provided carbon source. **b**, Schematic of porphyran PUL from NB001 based on alignment to the previously published *B. plebeius* PUL, based on data from whole-genome sequencing. Grey bar, the region deleted via homologous recombination to abolish the ability to utilize porphyran. **c**, NB001 has the ability to grow on porphyran (doubling time = 98 min) as the only carbon source (WT), and growth is abrogated when genes required for porphyran utilization are knocked out (KO).



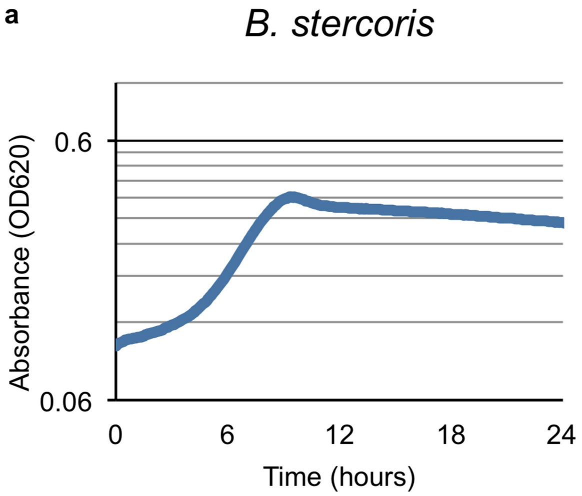
Extended Data Fig. 3 | Porphyran does not significantly impact the gut microbiota in the absence of a known utilizer. Weighted UniFrac analysis was performed on faecal 16S rRNA data for conventional mice colonized with a porphyran utilization knockout (as in Fig. 2c) before (Pre, $n = 8$) or after (Post, $n = 9$) addition of porphyran. **a**, Principal coordinates analysis. **b**, Weighted UniFrac analysis. Unpaired two-tailed t-test, $P = 0.25$ (n.s., not significant).



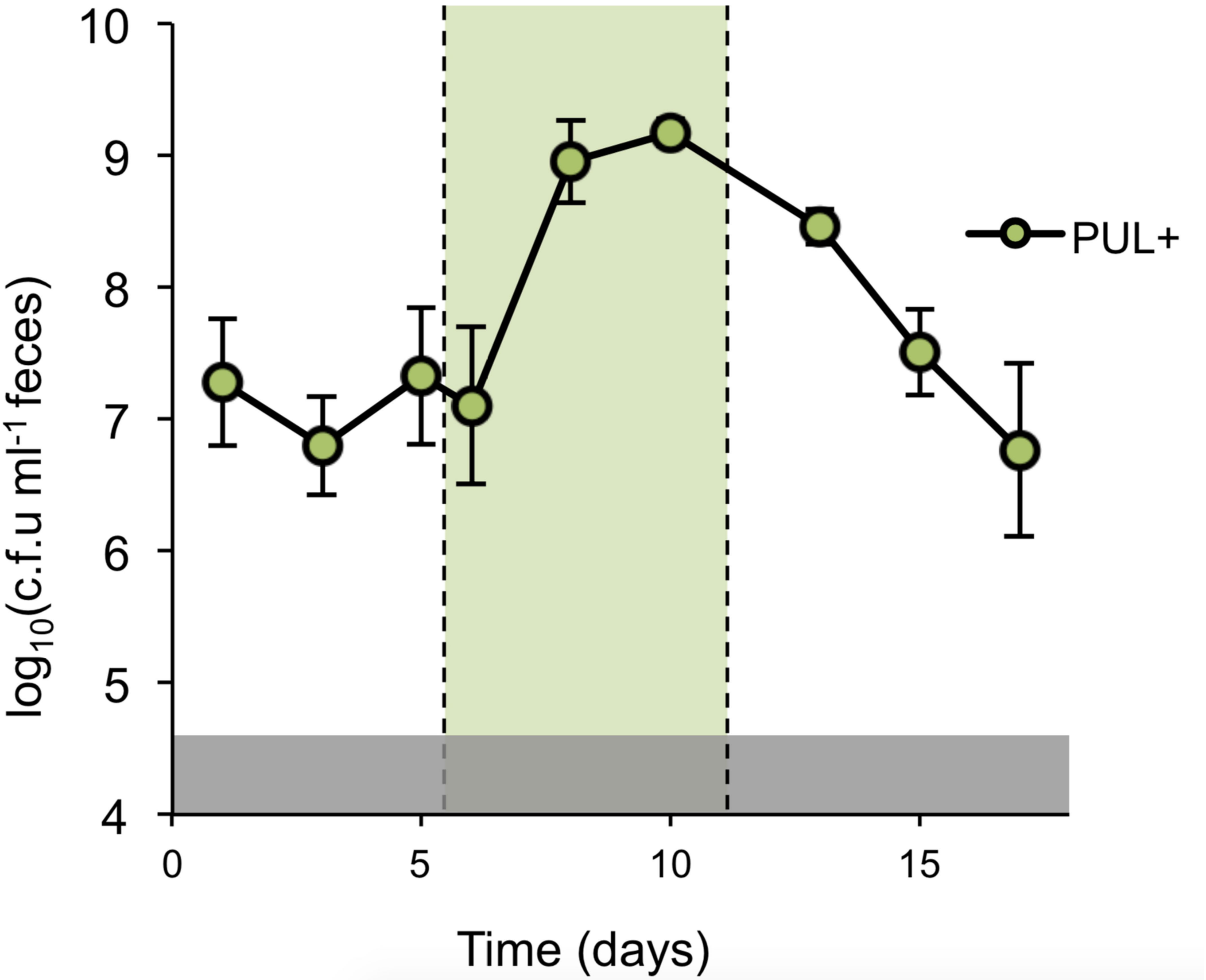
Extended Data Fig. 4 | Primary colonizer displacement is robust and contingent upon access to porphyran. a, Conventional mice ($n = 7$), which were fed a MAC-rich diet and colonized with NB001 (PUL^{-} 1) containing an eight-gene deletion abrogating its ability to utilize porphyran (Extended Data Fig. 2b, c), demonstrated resistance to subsequent challenge with an isogenic knockout strain (PUL^{-} 2) in the presence of 1% porphyran in the drinking water. Notably, our conventionally raised mice were permissive to colonization by this strain and other tested species of *Bacteroides* (*B. thetaiotaomicron*, *B. fragilis*, *B. uniformis*, *B. vulgatus*, stable colonization range of 8×10^5–3×10^8 c.f.u. per ml faeces), which differs from reports of tests on other conventionally raised mice, potentially reflecting inter-colony microbiota differences. **b**, Mice from Fig. 2e were challenged with the originally colonizing porphyran utilization knockout (PUL^{-}) that was displaced by the utilizer (PUL^{+}) and demonstrated colonization resistance to the previously displaced knockout strain. Data are mean ± s.d. The grey-shaded boxes represent the limit of detection.



Extended Data Fig. 5 | Minimal porphyran utilization PULs were constructed via PCR and yeast assembly. Schematic representing construction of designed porphyran PULs. On the basis of the alignment to the previously published *B. plebeius* porphyran utilization PUL, three regions were targeted for minimal PUL assembly and amplified via PCR from the NB001 genome. The PCR fragments were assembled with digests of both a custom and commercially available vector in yeast (see Methods), after which colonies carrying correctly assembled plasmids were lysed and directly added to *E. coli* for electroporation.



Extended Data Fig. 6 | *B. stercoris* and *B. thetaiotaomicron* demonstrate different abilities to grow in minimal medium. a, **b**, Wild-type *B. stercoris* (**a**) and *B. thetaiotaomicron* (**b**) grown in SMM with glucose as the only carbon source demonstrate different maximum optical densities reached (*B. stercoris* maximum OD = 0.363, *B. thetaiotaomicron* maximum OD = 0.453). This suggests a possible explanation for why both species with the 21-gene PUL grow to different maximum optical densities as well (Fig. 3b).



Extended Data Fig. 7 | Abundance of *B. stercoris* with the 34-gene PUL can be controlled in the context of a conventional mouse microbiota. Conventional mice ($n = 5$) that were fed a MAC-rich diet were colonized with a strain of *B. stercoris* harbouring the designed 34-gene porphyran PUL and density of the engineered strain was tracked in the faeces. Upon administration of 1% porphyran in the drinking water (green-shaded box), density of *B. stercoris* increased, and subsequently decreased upon removal of porphyran. Data are mean ± s.d. The grey-shaded box represents the limit of detection.

Extended Data Table 1 | List of plasmids used in this work

Plasmid Name	Description	Length (bps)	Parts used in yeast assembly	Relevant Figure
pWW3452	High GFP, erythromycin resistance	6,031	N/A	Fig. 1, Fig. 3e
pNBU2-ermGb	Erythromycin resistance	4,909	N/A	Fig. 2, Extended Data Fig. 3
pNBU2-tetQb	Tetracycline resistance	6,662	N/A	Fig. 2, Extended Data Fig. 3
pWD034	Porphyran PUL Deletion (main operon)	7,148	N/A	Fig. 2c
pWD035	Long Porphyran PUL	72,558	PCR_01, PCR_02, PCR_03, PCR_04, PCR_05, PCR_06, PCR_07, PCR_08, PCR_09, PCR_10, Vector_01, Vector_02, Vector_03	Fig. 3a, b
pWD036	Medium Porphyran PUL	53,695	PCR_04, PCR_05, PCR_06, PCR_07, PCR_08, PCR_10, PCR_11, Vector_01, Vector_02, Vector_03	Fig. 3a, b, c
pWD037	Short Porphyran PUL	32,364	PCR_05, PCR_06, PCR_12, PCR_13, Vector_01, Vector_02, Vector_03	Fig 3a, b
cLS428	High mCherry, tetracycline resistance	7,621	N/A	Fig. 4

N/A, not applicable.

LETTER

https://doi.org/10.1038/s41586-018-0110-6

Control of cardiac jelly dynamics by NOTCH1 and NRG1 defines the building plan for trabeculation

Gonzalo del Monte-Nieto[1,2]*, Mirana Ramialison[3], Arne A. S. Adam[1], Bingruo Wu[4], Alla Aharonov[5], Gabriele D'Uva[5], Lauren M. Bourke[6,7], Mara E. Pitulescu[8,9], Hanying Chen[10], José Luis de la Pompa[11,12], Weinian Shou[10], Ralf H. Adams[8,9], Sarah K. Harten[6,7], Eldad Tzahor[5], Bin Zhou[4] & Richard P. Harvey[1,2,13]*

In vertebrate hearts, the ventricular trabecular myocardium develops as a sponge-like network of cardiomyocytes that is critical for contraction and conduction, ventricular septation, papillary muscle formation and wall thickening through the process of compaction[1]. Defective trabeculation leads to embryonic lethality[2–4] or non-compaction cardiomyopathy (NCC)[5]. There are divergent views on when and how trabeculation is initiated in different species. In zebrafish, trabecular cardiomyocytes extrude from compact myocardium[6], whereas in chicks, chamber wall thickening occurs before overt trabeculation[7]. In mice, the onset of trabeculation has not been described, but is proposed to begin at embryonic day 9.0, when cardiomyocytes form radially oriented ribs[2]. Endocardium–myocardium communication is essential for trabeculation, and numerous signalling pathways have been identified, including Notch[2,8] and Neuregulin (NRG)[4]. Late disruption of the Notch pathway causes NCC[5]. Whereas it has been shown that mutations in the extracellular matrix (ECM) genes *Has2* and *Vcan* prevent the formation of trabeculae in mice[9,10] and the matrix metalloprotease ADAMTS1 promotes trabecular termination[3], the pathways involved in ECM dynamics and the molecular regulation of trabeculation during its early phases remain unexplored. Here we present a model of trabeculation in mice that integrates dynamic endocardial and myocardial cell behaviours and ECM remodelling, and reveal new epistatic relationships between the involved signalling pathways. NOTCH1 signalling promotes ECM degradation during the formation of endocardial projections that are critical for individualization of trabecular units, whereas NRG1 promotes myocardial ECM synthesis, which is necessary for trabecular rearrangement and growth. These systems interconnect through NRG1 control of *Vegfa*, but act antagonistically to establish trabecular architecture. These insights enabled the prediction of persistent ECM and cardiomyocyte growth in a mouse NCC model, providing new insights into the pathophysiology of congenital heart disease.

Here we analysed the two- and three-dimensional features of trabeculation in the left ventricle from heart tube formation (embryonic day (E)8.0) until the beginning of compaction (E14.5) and define the phases of sprouting, touchdown formation, lateral ingression, extension and termination (Fig. 1a–g, Extended Data Fig. 1a–g and Supplementary Figs. 1–7). Sprouting begins at E8.0, when the endocardium forms sprouts that are oriented towards the myocardium. The myocardial wall is already multilayered at this time with discontinuous luminal cardiomyocyte clusters (laminar trabeculae) that are loosely attached to the compact layer, suggesting that trabeculation has begun (Fig. 1a and Extended Data Fig. 1a). Touchdowns occur at E8.0–E8.5 as endocardial projections progress through cardiac jelly, form distinct columns and establish the first anchor points with compact myocardium exclusively at the ventricular outer curvature. Endocardial ridges then form between touchdowns, creating endocardial domes that segment the ventricular chamber into multiple matrix-rich areas (ECM bubbles), thus encapsulating the trabecular myocardium and defining the initial trabecular architectural units (Fig. 1b, c, Extended Data Fig. 1b, c, e, f, h–j and Supplementary Figs. 1–7). Lateral ingression occurs from E8.5, as endocardial touchdowns expand laterally beneath laminar trabeculae, further isolating trabecular units (Fig. 1c and Extended Data Fig. 1c). Trabecular assembly occurs at E8.5–E9.0 as trabecular cardiomyocytes rearrange into a radial disposition (Fig. 1d and Extended Data Fig. 1e, f). Finally, trabecular extension occurs from E9.5 when trabeculae undergo radially oriented cardiomyocyte growth and ECM bubbles are progressively reduced in a basal-to-apical manner. The peak of trabecular growth occurs around E14.5 (termination)[3], when ECM is minimalized, whereas trabecular simplification and further ECM reduction occurs afterwards during compaction (Fig. 1e–g and Extended Data Fig. 1k–m).

The trabecular markers *Hopx*, *Gja5* and *Bmp10* were not enriched in laminar trabeculae at E8.0 (Extended Data Fig. 1n–p). However, the ECM synthesis genes *Has2* and *Vcan*, which are essential for trabeculation[9,10], and CD44 (a hyaluronan receptor), were enriched in laminar trabeculae (E8.0–E9.0), became progressively restricted to the trabecular apex, which is associated with ECM bubbles (E9.5–E13.5), and disappeared after termination (E16.5) (Extended Data Fig. 2a, e, g). Immunofluorescence analysis showed that ECM components were deposited adjacent to both the endocardium and myocardium, even when these components were mainly synthesized by the myocardium (Extended Data Fig. 2a, c). Among the ECM components that were analysed (Extended Data Fig. 2c, h–m), hyaluronic acid (detected with hyaluronic acid-binding protein; HABP) and fibronectin showed patterns that were similar to those found for *Has2*, *Vcan* and CD44 (Extended Data Fig. 2c, h). By contrast, ECM proteolytic genes *Adamts1*, *Mmp2* and *Hyal2* were enriched in touchdown endocardium at E8.0–E9.0. These genes later expanded to follow the basal-to-apical pattern of ECM degradation described above (E9.5–E11.5) and eventually occupied the entire endocardium (E13.5–E16.5) (Extended Data Fig. 3a–c). Neoversican, a readout of metalloprotease activity[3], showed more precisely the dynamics of ECM degradation (Extended Data Fig. 3d). Therefore, our data suggest that trabeculation involves dynamic ECM remodelling, and the myocardium and endocardium are engaged in ECM synthesis and degradation, respectively.

---

[1]Developmental and Stem Cell Biology Division, Victor Chang Cardiac Research Institute, Darlinghurst, New South Wales, Australia. [2]St Vincent's Clinical School, University of New South Wales, Kensington, New South Wales, Australia. [3]Australian Regenerative Medicine Institute, Monash University, Clayton, Victoria, Australia. [4]Departments of Genetics, Albert Einstein College of Medicine of Yeshiva University, Bronx, New York, NY, USA. [5]Department of Molecular Cell Biology, Weizmann Institute of Science, Rehovot, Israel. [6]Epigenetics Laboratory, QIMR Berghofer Medical Research Institute, Herston, Queensland, Australia. [7]School of Biomedical Sciences, Institute of Health and Biomedical Innovation, Queensland University of Technology, Brisbane, Queensland, Australia. [8]Department of Tissue Morphogenesis, Max Planck Institute for Molecular Biomedicine, Münster, Germany. [9]Faculty of Medicine, University of Münster, Münster, Germany. [10]Departments of Pediatrics and Medical and Molecular Genetics, Riley Heart Research Center, Wells Center for Pediatric Research, Indiana University, Indianapolis, IN, USA. [11]Intercellular Signalling in Cardiovascular Development and Disease Laboratory, Centro Nacional de Investigaciones Cardiovasculares Carlos III (CNIC), Madrid, Spain. [12]Ciber cardiovascular, ISCIII, Madrid, Spain. [13]School of Biotechnology and Biomolecular Science, University of New South Wales, Kensington, New South Wales, Australia. *e-mail: g.delmonte@victorchang.edu.au; r.harvey@victorchang.edu.au

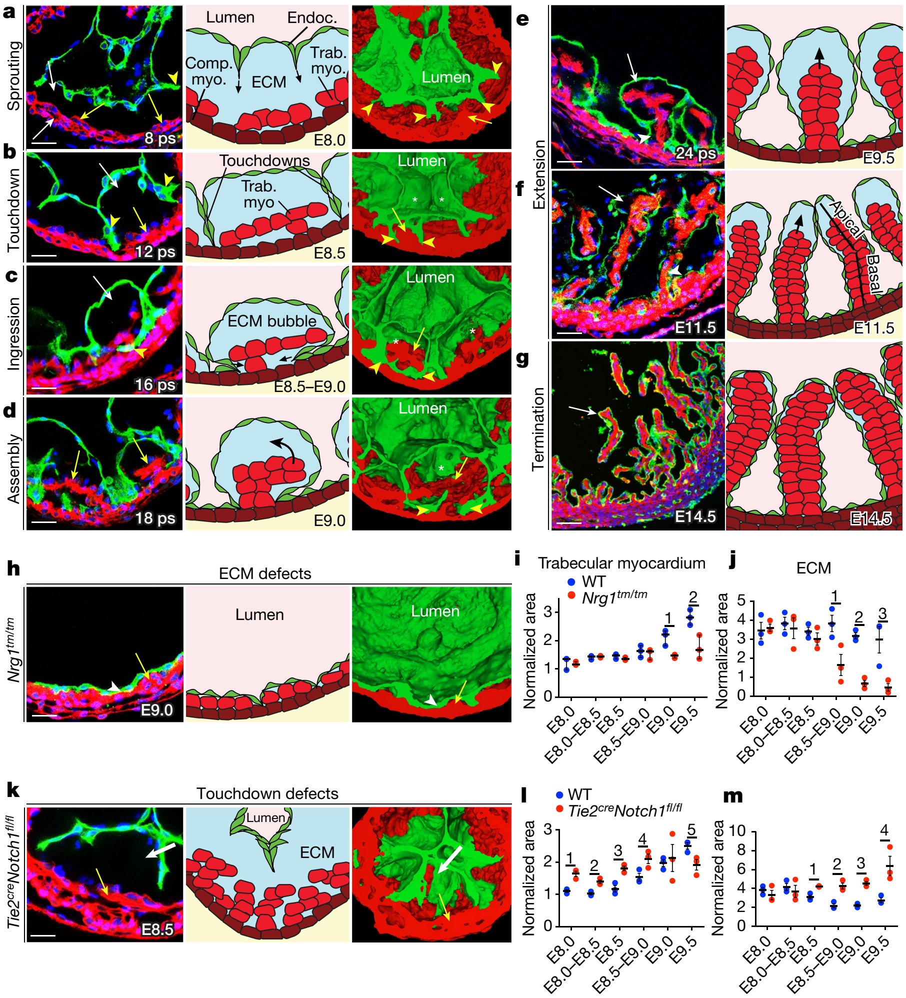


Fig. 1 | Cardiac trabeculation stages. **a**–**h**, **k**, Left, optical sections of left ventricles showing endocardium (GFP, green) and other tissues (tdTomato, red) of wild-type (WT; **a**–**g**), $Nrg1^{tm/tm}$ (**h**) and $Tie2^{cre}Notch1^{fl/fl}$ (**k**) embryos carrying $Tie2^{cre}R26^{mTmG}$. Yellow arrowheads, endocardium; yellow arrows, trabecular myocardium; white arrows, ECM. **a**–**h**, **k**, Middle (**a**–**d**, **h**, **k**) and right (**e**–**g**), cartoons representing lumen (pink), endocardium (endoc., green), cardiac jelly (ECM bubble, blue), trabecular myocardium (trab. myo., red), compact myocardium (comp. myo., brown) in wild-type (**a**–**g**), $Nrg1^{tm/tm}$ (**h**) and $Tie2^{cre}Notch1^{fl/fl}$ (**k**) embryos. Arrows indicate tissue dynamics. **a**–**d**, **h**, **k**, Right, luminal views of transected 3D reconstructions in wild-type (**a**–**d**), $Nrg1^{tm/tm}$ (**h**) and $Tie2^{cre}Notch1^{fl/fl}$ (**k**) embryos. Green, endocardium; red, myocardium; void between endocardium and myocardium, ECM. Yellow arrowheads, endocardial sprouts/touchdowns; yellow arrows, trabecular myocardium; white asterisks, ECM bubbles/endocardial domes. See Supplementary Figs. 1–7. Scale bars, 20 μm (**a**, **c**, **f**), 15 μm (**b**), 25 μm (**d**, **e**, **h**, **k**) and 60 μm (**g**). **i**, **j**, **l**, **m**, Quantification of trabecular myocardium (**i**, **l**) and ECM (**j**, **m**) areas in $Nrg1^{tm/tm}$ (**i**, **j**) and $Tie2^{cre}Notch1^{fl/fl}$ (**l**, **m**) versus wild-type embryos. Data are mean ± s.e.m. Two-sided Student's *t*-tests (without corrections for multiple comparisons) were used. Significant comparisons are indicated by numbers as follows. **i**, 1, $P = 0.014$; 2, $P = 0.021$. **j**, 1, $P = 0.036$; 2, $P = 0.00039$; 3, $P = 0.0258$. **l**, 1, $P = 0.004$; 2, $P = 0.0047$; 3, $P = 0.0064$; 4, $P = 0.04$; 5, $P = 0.038$. **m**, 1, $P = 0.0056$; 2, $P = 0.0057$; 3, $P = 0.00058$; 4, $P = 0.026$. **a**–**m**, $n = 3$ independent embryos per genotype.

Among known mutant mouse lines with trabeculation defects, two categories of interest—those with premature loss of ECM and those with touchdowns defects—were identified. We defined the phenotypic onset stage for each mutant to avoid secondary network changes (Methods). Premature loss of ECM and reduced trabecular myocardium were features of mice with mutations in *Nrg1* (Fig. 1h–j, Extended Data Fig. 1f, g and Supplementary Figs. 1–7) and the NRG1 receptor gene *Erbb2* (Extended Data Fig. 4a, b). On the basis of published data, mutations in *Vcan*[10], *Has2*[9], *Brg1* (also known as *Smarca4*)[3] and cerebral cavernous malformation genes[11] could also be classified in this group. The second category, featuring touchdown simplification, excessive ECM and disorganized trabeculae, included $Tie2^{cre}Notch1^{fl/fl}$ (Fig. 1k–m and Extended Data Fig. 1c, d), $Tie1^{cre}VegfR2^{fl/fl}$ (Extended Data Fig. 4g, h) and $Tie2^{cre}Rbpj^{fl/fl}$ mouse embryos (Extended Data Fig. 4k, l). *Efnb2* and *Bmp10*, although implicated in trabeculation[2], were not essential for these early processes (Extended Data Fig. 4m–p). Our analysis assigned *Notch1* and *Nrg1* mutations to opposite categories, in contrast to a previous study[2], prompting us to perform further analyses.

We first investigated the morphology of the ventricular chambers in the mutant mouse embryos. In $Nrg1^{tm/tm}$ embryos[4], the early phases of trabeculation occurred normally (Extended Data Figs. 1q, 4q, r). However, from E8.5 to E9.0, there was a rapid reduction in ECM, promoting 'compression' of the laminar trabeculae, and preventing trabecular rearrangement and expansion (Fig. 1h-j and Extended Data Fig. 1f, g, 4s–u). Ventricular chambers of $Tie2^{cre}Notch1^{fl/fl}$ embryos at E8.0–E9.0 showed reduced touchdowns, absence of endocardial domes, excessive ECM accumulation, reduced lumen and disorganized laminar trabeculae (Fig. 1k–m and Extended Data Fig. 1c, d, r, 4q–t). Unlike other touchdown defect mutants, $Tie2^{cre}Notch1^{fl/fl}$ mice showed a significant increase in trabecular myocardium in early phases (E8.0–E9.0; Fig. 1k, l), although this was reduced at later stages, close to lethality (E9.5; Fig. 1l and Extended Data Fig. 4u). Only the reduction in trabecular myocardium has been described previously[2]. Our data suggest roles for NRG1 in ECM bubble maintenance and for NOTCH1 in endocardial touchdown/dome formation, and that both processes are essential for trabecular organization and growth.

We then investigated the expression of genes associated with the development of the ventricular chambers. In $Nrg1^{tm/tm}$ hearts at E9.0, expression levels of *Nppa*, *Mest*, *Hopx*, *Gja5*, *Mycn* and *Hand2* were unchanged, whereas *Bmp10* and *Hey2* were ectopically expressed throughout the myocardium (Extended Data Fig. 5a–i), suggesting defective separation of trabecular and compact layer identities. However, the lack of an ECM phenotype in *Bmp10* mutant mice (Extended Data Fig. 4o, p) and the late (perinatal) phenotype in *Hey2* mutant mice[12] suggest that these genes are not involved in early ECM dynamics. In $Tie2^{cre}Notch1^{fl/fl}$ mice at E8.5, expression of chamber markers was normal, including *Hand2* and *Bmp10*, which have been reported to be dependent on Notch signalling[2,13] (Extended Data Fig. 5a–i). However, at later stages (E9.5), ventricular chamber genes showed changes in both strains as has been reported previously[2,4,13] (Extended Data Fig. 5i). Thus, defects in mutant mice at the phenotypic onset stage appear unrelated to changes in chamber gene expression.

We subsequently assessed ECM synthesis and degradation. In $Nrg1^{tm/tm}$ embryos at E9.0, expression of *Has2*, *Vcan*, HABP and CD44 was mostly lost in the trabecular myocardium, but was upregulated and ectopically expressed throughout the myocardium in $Tie2^{cre}Notch1^{fl/fl}$ embryos at E8.5 (Fig. 2a, b and Extended Data Fig. 6a–h, l). Other ECM synthesis markers showed no changes, consistent with the lack of

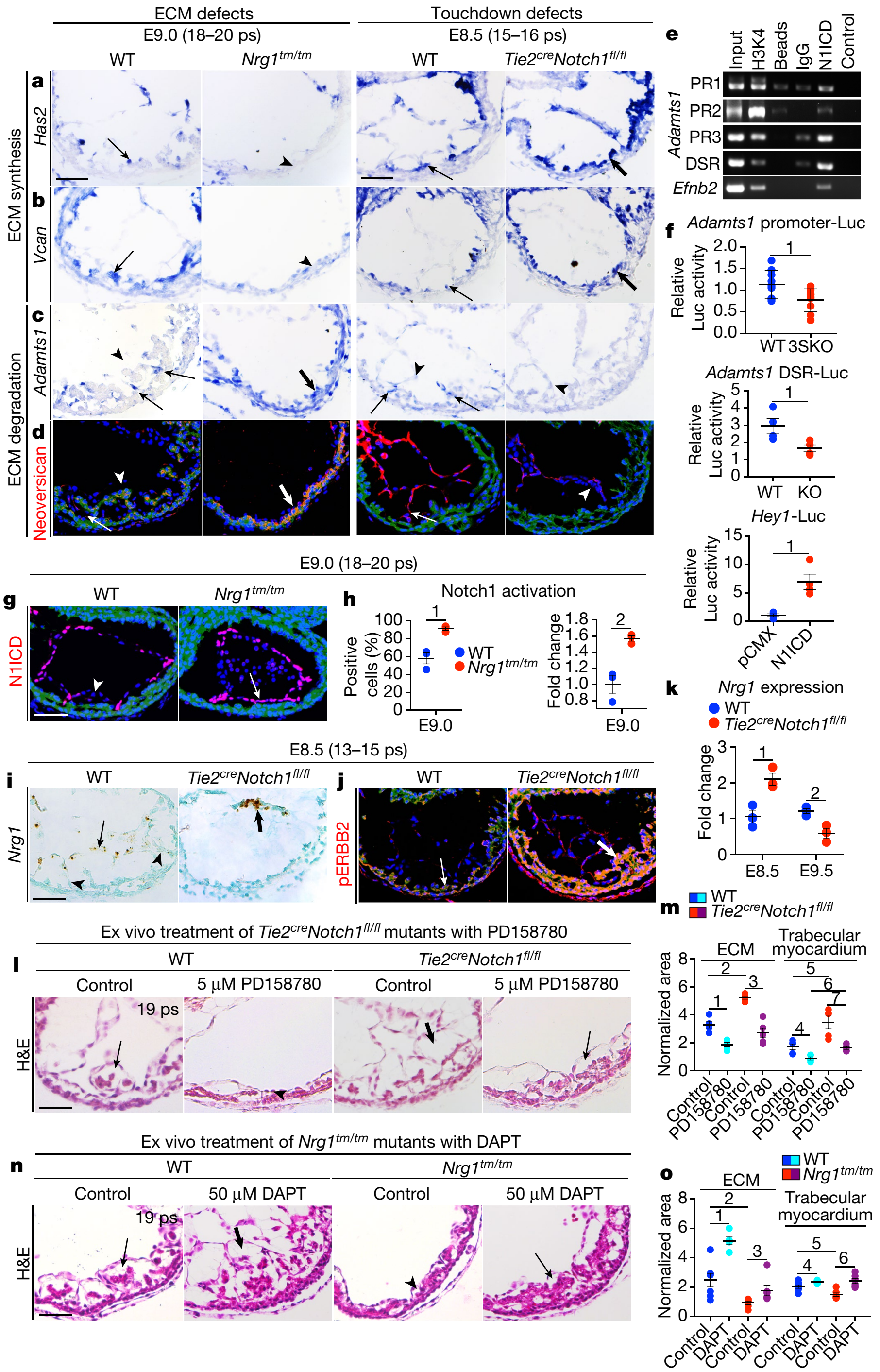


Fig. 2 | NRG1 and NOTCH1 pathways promote ECM synthesis and degradation, respectively. a–d, g, i–k, Left ventricles of *Nrg1*$^{tm/tm}$ (E9.0) and *Tie2*^{cre}*Notch1*$^{fl/fl}$ (E8.5) embryos were compared to wild-type embryos. The indicated markers were analysed by in situ hybridization (**a–c**), immunofluorescence (**d, g, j**), RNA-scope (**i**) and qPCR (**k**). Arrows, expression; arrowheads, reduced expression; thick arrows, increased expression. In immunofluorescence panels, the myocardium is stained for SMA (green dulled mask) and with DAPI (nuclei, blue). **h**, Percentage of and fold change in the number of N1ICD$^+$-expressing cells. **k**, *Nrg1* expression levels in *Notch1* mutants at E8.5 and E9.5. **e**, N1ICD chromatin immunoprecipitation followed by PCR for conserved RBPj-binding sites in predicted *Adamts1* promoter and enhancers in mouse aortic endothelial cells. *Efnb2* was used as a positive control. PR, promoter region (regions 1–3); DSR, downstream region. **f**, Luciferase activity assays for *Adamts1* promoter (top; 3SKO, 3 binding sites mutated), downstream enhancer (middle; KO, binding site mutated) and *Hey1*-Luc positive control (bottom; pCMX, empty vector). **l–o**, Morphology of *Tie2*^{cre}*Notch1*$^{fl/fl}$ embryos treated with 5 μM PD158780 or DMSO control (**l, m**) and *Nrg1*$^{tm/tm}$ embryos treated with 50 μM DAPT or ethanol control (**n, o**), compared to wild-type in RBECs for 24 h. **m, o**, Quantification of the ECM and trabecular myocardium area. **a–d, h–j**, $n = 3$ independent embryos per genotype; **m, o**, $n = 5$ in except in **o** $n = 7$ for both controls and $n = 6$ for *Nrg1*$^{tm/tm}$ treated with DAPT; **f**, $n = 9$ independent repeats (top); $n = 4$ (middle, bottom). **k**, $n = 3$ biological replicates with three experimental replicates. Data are mean ± s.e.m. Two-sided Student's *t*-tests (without corrections for multiple comparisons) were used. Significant comparisons are indicated by numbers as follows. **f**, 1, $P = 0.019$; 2, $P = 0.049$; 3, $P = 0.02$. **h**, 1, $P = 0.007$; 2, $P = 0.007$. **k**, 1, $P = 0.013$; 2, $P = 0.015$. **m**, 1, $P = 0.0007$; 2, $P = 0.000045$; 3, $P = 0.00013$; 4, $P = 0.007$; 5, $P = 0.0075$; 6, $P = 0.0008$; 7, $P = 0.0041$. **o**, 1, $P = 0.001$; 2, $P = 0.0096$; 3, $P = 0.049$; 4, $P = 0.045$; 5, $P = 0.008$; 6, $P = 0.0013$. Scale bars, 60 μm (**a–d** (left two panels), **g, l, n**) and 50 μm (**a–d** (right two panels), **i, j**).

early cardiac phenotypes in cognate mutant mice. Time-course analysis (Extended Data Figs. 2a–f, 4q) showed that NRG1 signalling was not required for cardiac jelly synthesis before E8.0, but was essential for its subsequent maintenance and compartmentalization. In *Nrg1*$^{tm/tm}$ embryos at E9.0, *Adamts1*, *Mmp2* and neoversican were ectopically expressed in the endocardium and myocardium, however, these genes were downregulated in the endocardium of *Tie2*^{cre}*Notch1*$^{fl/fl}$ embryos at E8.5, whereas other ECM degradation markers were unchanged (Fig. 2c, d and Extended Data Fig. 6i–l). BRG1, a repressor of *Adamts1* in the endocardium[3], was dysregulated in the myocardium of *Nrg1*$^{tm/tm}$ mice (around 33% reduction in positive cells), whereas no changes were seen in the endocardium or in *Tie2*^{cre}*Notch1*$^{fl/fl}$ embryos (Extended Data Fig. 6m, n). Thus, myocardial BRG1 downregulation in *Nrg1*$^{tm/tm}$ embryos may explain *Adamts1* ectopic activation in this tissue. We then investigated whether changes in proliferation occurred. No changes in chamber myocardium proliferation were observed in mutants at phenotypic onset stages. However, we found a decrease in the proportion of trabecular cardiomyocytes in *Nrg1*$^{tm/tm}$ mice, and an increase in *Tie2*^{cre}*Notch1*$^{fl/fl}$ embryos at the phenotypic onset and later stages (Extended Data Fig. 6r, s). Proliferation at later stages was reduced in both mutants as has been described previously[2,4] (Extended Data Fig. 6t, u). Collectively, our data suggest that the *Nrg1*$^{tm/tm}$ phenotype results from both reduced ECM synthesis and excessive ECM degradation. Conversely, the *Tie2*^{cre}*Notch1*$^{fl/fl}$ phenotype arises from both a lack of ECM degradation and ectopic ECM synthesis. Reciprocal changes in the trabecular proportion suggest a role for these pathways in cardiomyocyte allocation to the trabecular layer.

Gene expression analysis using RNA-scope showed that *Nrg1* expression is enriched in the endocardium away from touchdowns at E8.0–E9.0, and becomes progressively restricted to ECM bubble

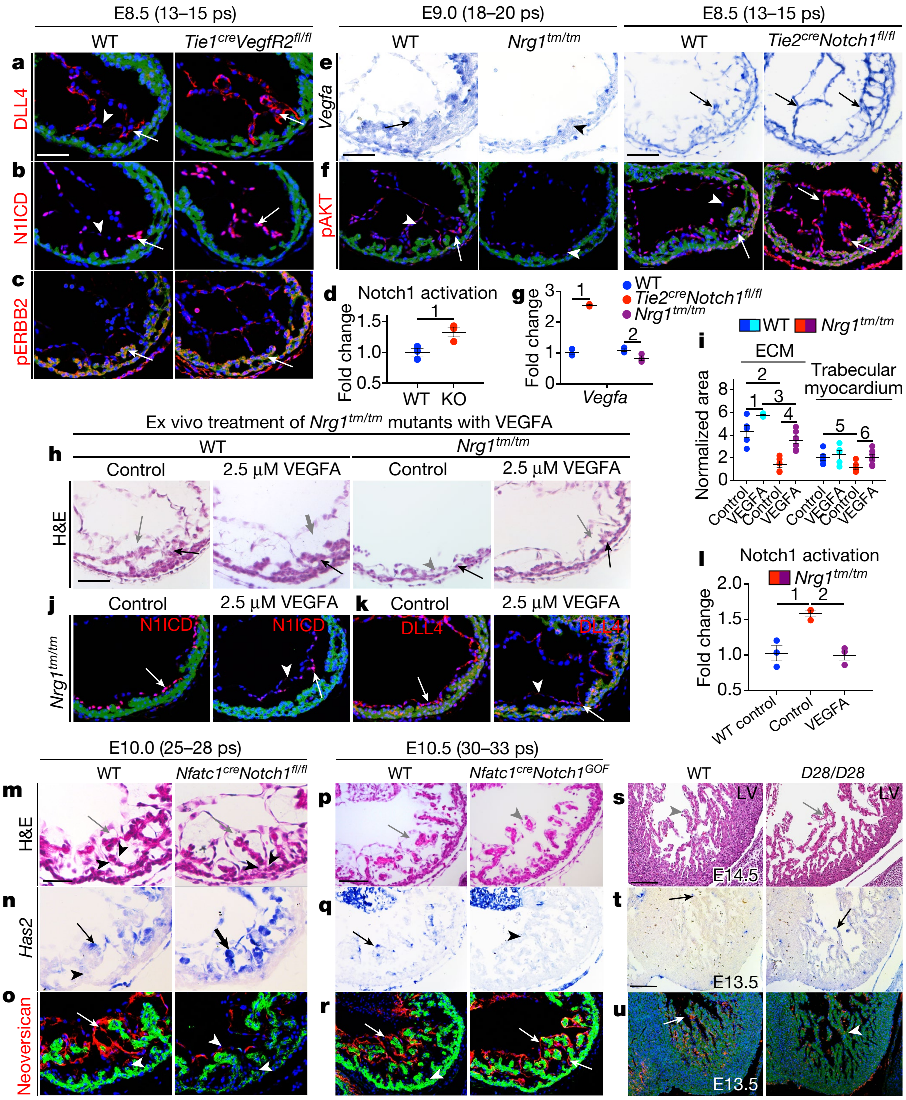


Fig. 3 | NRG1-dependent VEGFA signalling promotes N1ICD restriction, late trabeculation phases and implications of the model for NCC. a–u, Morphological and molecular analysis in left ventricle (LV) of *Tie1*^{cre}*Vegfr2*$^{fl/fl}$ (E8.5), *Nrg1*$^{tm/tm}$ (E9.0), *Tie2*^{cre}*Notch1*$^{fl/fl}$ (E8.5), *Nfatc1*^{cre}*Notch1*$^{fl/fl}$ (E10.0), *Nfatc1*^{cre}*Notch1*^{GOF} (E10.5) and *Rlf*$^{D28/D28}$ (E13.5, E14.5) mutant embryos compared to wild-type. **h–k**, *Nrg1*$^{tm/tm}$ embryos treated with 2.5 μM VEGFA or PBS control in RBECs for 24 h. Analysis was carried out using immunofluorescence (**a–c**, **f**, **j**, **k**, **o**, **r**, **u**), in situ hybridization (**e**, **n**, **q**, **t**), haematoxylin and eosin (H&E; **h**, **m**, **p**, **s**) and qPCR (**g**) using indicated markers. for gene and protein expression: arrows, expression; arrowheads, reduced expression; thick arrows, increased expression. In haematoxylin and eosin panels: black arrows, trabecular myocardium; grey arrows, ECM; black arrowheads, endocardium; grey arrowheads, reduced ECM; thick grey arrows, increased ECM. In immunofluorescence panels, the myocardium is stained for SMA (green dulled mask) and for nuclei with DAPI (blue). **d**, **l**, Fold change in number of N1ICD^{+} cells. **i**, ECM and trabecular myocardium area quantifications. **a–f**, **j–r**, **t**, **u**, $n = 3$ independent embryos per genotype; **s**, $n = 4$ (left) and $n = 6$ (right); **i**, **h**, $n = 5$, except for $n = 7$ *Nrg1*$^{tm/tm}$ embryos treated with VEGFA. **g**, n = 3 biological replicates with three experimental replicates. Data are mean ± s.e.m. Two-sided Student's *t*-tests (without corrections for multiple comparisons) were used. Significant comparisons are indicated by numbers as follows. **d**, 1, $P = 0.032$. **g**, 1, $P < 0.0001$; 2, $P = 0.04$. **i**, 1, $P = 0.026$; 2, $P = 0.0013$; 3, $P = 0.00011$; 4, $P = 0.0006$; 5, $P = 0.034$; 6, $P = 0.027$. l, 1, $P = 0.02$; 2, $P = 0.0039$. Scale bars: 50 μm (**a–c**, **e**, **f** (right two panels)), 55 μm (**e**, **f** (left two panels), **h**, **j**, **k**), 60 μm (**m–r**), 85 μm (**t**, **u**[1]) and 150 μm (**s**).

endocardium (E9.5–E13.5), and finally to the trabecular apex (E16.5) (Extended Data Fig. 3e). Active ERBB2 (phosphorylated (p)ERBB2, a NRG1 signalling readout; Extended Data Fig. 4e, f) was detected in both trabecular and compact cardiomyocytes at E8.0, but became enriched in trabecular myocardium associated with ECM bubbles at E8.5–E16.5 (Extended Data Fig. 3g), correlating with ECM synthesis gene expression (Extended Data Fig. 2a, e, g). These results, together with the downregulation of ECM synthesis genes in *Nrg1*$^{tm/tm}$ embryos, suggest that NRG1 signalling promotes ECM synthesis. Conversely, patterns of expression of ECM degradation genes in the endocardium (Extended Data Fig. 3a–c) mirrored those of activated NOTCH1 (NOTCH1 intracellular domain; (N1ICD))[2] and the Notch ligand delta-like 4 (DLL4) (Extended Data Fig. 3i, k), suggesting that Notch promotes ECM degradation. Three conserved RBPj-binding sites within the *Adamts1* promoter region and one within a predicted downstream enhancer region showed N1ICD binding by chromatin immunoprecipitation (Fig. 2e), and specifically responded to N1ICD in luciferase reporter assays (Fig. 2f), confirming direct regulation of *Adamts1* by N1ICD.

We next tuned ECM dynamics by treating E8.0 wild-type and *Tie2*^{cre}*Notch1*$^{fl/fl}$ embryos in rolling bottle embryo cultures (RBECs) with recombinant hyaluronidase (HYAL), which promotes hyaluronic acid degradation, or wild-type and *Nrg1*$^{tm/tm}$ embryos with the metalloprotease inhibitor GM6001[3]. HYAL-treated wild-type embryos recapitulated the *Has2*[9] and *Nrg1*$^{tm/tm}$ mutant phenotypes, highlighting the importance of excessive ECM degradation in *Nrg1*$^{tm/tm}$ embryos (Extended Data Fig. 7a, b). HYAL-treated *Tie2*^{cre}*Notch1*$^{fl/fl}$ embryos also showed a reduction in ECM, but a complete reduction was delayed (Extended Data Fig. 7a, white arrows), because of the excessive ECM synthesis that is characteristic in this mutant (Fig. 2a, b). Ventricles of *Tie2*^{cre}*Notch1*$^{fl/fl}$ embryos already showed an increased trabecular myocardium area at the onset of RBEC culture (E8.0, Fig. 1l) due to excessive allocation of cardiomyocytes to the trabecular layer (Extended Data Fig. 6l). Slowed ECM degradation after HYAL treatment allowed continued trabecular myocardium formation as in untreated *Tie2*^{cre}*Notch1*$^{fl/fl}$ embryos, but trabeculae were eventually compacted (Extended Data Fig. 7a, b). GM6001-treated wild-type embryos showed reduced touchdowns, disorganized trabeculae and increased ECM (Extended Data Fig. 7c, d), resembling abnormalities in touchdown defect mutants and highlighting the importance of ECM degradation for touchdown formation and trabecular organization (Fig. 1k–m and Extended Data Fig. 4g, h, k, l). GM6001-treated *Nrg1*$^{tm/tm}$ embryos showed pronounced rescue of ECM, decompression of laminar trabeculae and reduced touchdowns (Extended Data Fig. 7c, d), although cardiomyocyte expansion did not occur, probably because correctly regulated ECM dynamics is needed for trabecular growth. These results confirm that correct ECM dynamics are necessary for endocardial touchdown and ECM bubble formation, and critical for individualization and organization of trabecular units.

We then explored NOTCH1 pathway activity in NRG1 pathway mutants. NOTCH1 was ectopically activated in *Nrg1*$^{tm/tm}$ and *Erbb2*$^{-/-}$ mutants throughout the ventricular endocardium before ECM

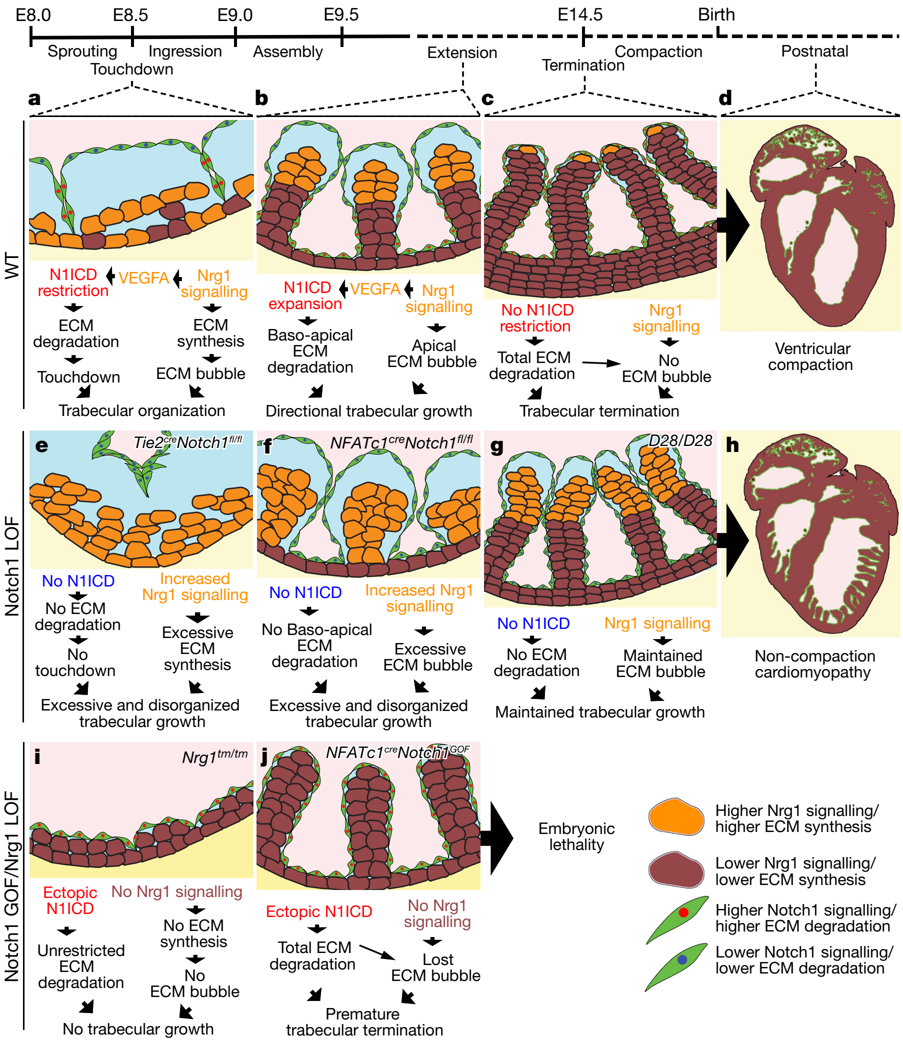


Fig. 4 | Models of normal and abnormal trabeculation. a–j, Timeline, cartoon morphological elements and pathway changes are shown. **a–d**, features of the model of cardiac trabeculation in wild-type left ventricle. **e–h**, features for early *Tie2*^{cre}*Notch1*$^{fl/fl}$ (**e**), mid-gestation *Nfatc1*^{cre}*Notch1*$^{fl/fl}$ (**f**) and late *Rlf* (**g**, **h**) mutants. **i**, **j**, features of *Nrg1*$^{tm/tm}$ (**i**) and *Nfatc1*^{cre}*Notch1*^{GOF} (**j**) mutants.

reduction (90% of cells; 1.57-fold increase; Fig. 2g, h and Extended Data Fig. 3i, j, 4c, d). DLL4 and *Efnb2* (regulated by NOTCH1[2]), were also ectopically expressed (Extended Data Fig. 6l, o, p). Notably, an endothelial layer-specific NOTCH1 gain-of-function (GOF) mutant (*Tie2*^{cre}*Notch1*^{GOF}) showed a phenotype similar to *Nrg1*$^{tm/tm}$ mutants from E9.5 (Extended Data Fig. 7f, g), with ectopic expression of *Adamts1*, *Mmp2* and neoversican in the endocardium (Extended Data Fig. 7h–j, m, n, q), confirming that ectopic Notch signalling promotes excessive ECM degradation. However, the degradation of ECM in *Tie2*^{cre}*Notch1*^{GOF} embryos was delayed compared to *Nrg1*$^{tm/tm}$ embryos because the expression of ECM synthesis elements (*Nrg1*, pERBB2, *Has2*, *Vcan)* was normal initially (Extended Data Fig. 7e, g, j–l), albeit downregulated by E9.5 (Extended Data Fig. 7o–q). These data confirm that the *Nrg1*$^{tm/tm}$ phenotype is more severe, as it involves the combination of excessive endocardial ECM degradation driven by ectopic Notch expression, as well as ectopic ECM degradation and reduction in ECM synthesis from the myocardium.

From E8.5, *Nrg1* and pERBB2 expression was markedly upregulated in *Tie2*^{cre}*Notch1*$^{fl/fl}$ endocardium and myocardium, respectively (Fig. 2i–k and Extended Data Fig. 3e–h), suggesting that in these mutants excessive NRG1 activity contributes to excessive ECM synthesis. However, *Nrg1* and pERBB2 were downregulated closer to lethality (E9.5) as has been described previously[2] (Fig. 2k and Extended Data Fig. 3e–h). Overall, these results confirm the regulation of ECM synthesis and degradation by NRG1 and NOTCH1 pathways, respectively, and their reciprocal dysregulation in mutant embryos.

To explore reciprocal interactions further, we treated wild-type and *Tie2*^{cre}*Notch1*$^{fl/fl}$ mutants in RBECs with PD158780 (an ERBB inhibitor) or wild-type and *Nrg1*$^{tm/tm}$ mutants with DAPT (a Notch inhibitor[2]). PD158780- and DAPT-treated wild-type embryos were strong phenocopies of *Nrg1*$^{tm/tm}$ and *Tie2*^{cre}*Notch1*$^{fl/fl}$ mutants, respectively (Fig. 2l–o). However, PD158780 treatment of *Tie2*^{cre}*Notch1*$^{fl/fl}$ and DAPT treatment of *Nrg1*$^{tm/tm}$ mouse embryos induced marked phenotypic rescue, with normalization of touchdown complexity, and trabecular myocardium and ECM areas, leading to trabeculae sequestered within ECM bubbles (Fig. 2l–o). Rescue also occurred in *Nrg1*$^{tm/tm}$ mice in which Notch signalling was inactivated genetically with *Tie2*^{cre}*Rbpj*$^{fl/fl}$ or *Tie2*^{cre}*Notch1*$^{fl/fl}$ alleles (Extended Data Fig. 8a–f). ECM that is present in double mutants probably reflects the persistence of original ECM (NRG1-independent; Extended Data Fig. 4q) due to blocked Notch-dependent ECM degradation. Touchdown formation could be explained by mosaicism in Cre-mediated deletion, allowing some N1ICD to become restricted to touchdown endocardium (Extended Data Fig. 8c, d, g). These results reinforce the reciprocal relationship between NOTCH1 and NRG1 signalling in ECM dynamics and cardiomyocyte allocation, as fundamental determinants of trabecular architecture.

We next treated E8.0 wild-type and *Nrg1*$^{tm/tm}$ embryos in RBECs with NRG1. Treated wild-type embryos showed excessive ECM (Extended Data Fig. 8h, i), as was found in *Tie2*^{cre}*Notch1*$^{fl/fl}$ mutants, consistent with NRG1 control of ECM synthesis. NRG1-treated *Nrg1*$^{tm/tm}$ embryos showed a marked morphological rescue, with trabeculae organized within ECM bubbles, and restoration of N1ICD and DLL4 restriction to touchdown endocardium (Extended Data Fig. 8h–l), suggesting that NRG1 signalling promotes restriction of NOTCH1 activity. In angiogenesis, vascular endothelial growth factor A (VEGFA) signals through VEGFA receptor 2 (VEGFR2) leading to DLL4–NOTCH1 pathway restriction[14], and NRG1 induces VEGFA[15]. *Vegfa* haploinsufficiency in mice leads to trabecular defects[16]. In wild-type embryos, *Vegfa* was expressed in the myocardium in a pattern resembling pERBB2 (Extended Data Fig. 3g, l), whereas VEGFR2 was expressed in the endocardium in a pattern resembling DLL4 and N1ICD (Extended Data Fig. 3i, k, m). We analysed endothelial layer-specific *Vegfr2* mutants (*Tie1*^{cre}*Vegfr2*$^{fl/fl}$; Extended Data Fig. 4g, h) after confirming complete loss of endocardial VEGFR2 and active AKT (pAKT; VEGFR2 signalling readout) (Extended Data Fig. 4i, j). At E8.5, ventricles from *Tie1*^{cre}*Vegfr2*$^{fl/fl}$ embryos showed a lack of DLL4 and N1ICD restriction in the endocardium, but they had normal myocardial NRG1 signalling (Fig. 3a–d). Therefore, VEGFA–VEGFR2 signalling is critical for Notch pathway restriction during trabeculation.

In *Nrg1*$^{tm/tm}$ mice, expression of *Vegfa* in the myocardium and pAKT in the endocardium were downregulated (Fig. 3e–g), indicating NRG1-dependent regulation of *Vegfa* in the myocardium. Similar to N1ICD and DLL4, endocardial VEGFR2 failed to regionalize (Extended Data Fig. 6q). Conversely, in *Tie2*^{cre}*Notch1*$^{fl/fl}$ mutants, *Vegfa*, VEGFR2, pAKT and DLL4 were markedly upregulated (Fig. 3e–g and Extended Data Fig. 6l, o, q), probably driven by upregulated NRG1 (Fig. 2i–k). To confirm the importance of proper VEGFA signalling in Notch pathway regionalization, we treated wild-type and *Nrg1*$^{tm/tm}$ embryos in RBECs with VEGFA. Treated wild-type embryos showed reduced touchdown complexity and increased ECM (Fig. 3h, i). VEGFA-treated *Nrg1*$^{tm/tm}$ embryos showed significant morphological rescue, including prevention of total matrix loss, restoration of trabeculation, and N1ICD and DLL4 restriction to touchdowns (Fig. 3h, i). We also treated *Tie2*^{cre}*Notch1*$^{fl/fl}$ mutants with SU5416 (a VEGFR2 inhibitor). SU5416-treated wild-type embryos were a precise phenocopy of *Tie1*^{cre}*VEGFR2*$^{fl/fl}$ mutants, including loss of restriction of N1ICD and DLL4 in the endocardium (Extended Data Fig. 4g, h, 8m–q). However, no morphological changes were observed in SU5416-treated *Tie2*^{cre}*Notch1*$^{fl/fl}$ embryos (Extended Data Fig. 8m, n), consistent with VEGF signalling acting upstream of NOTCH1. These results reveal that NRG1-dependent VEGFA signalling promotes restriction of NOTCH1 activity to touchdown endocardium.

We next manipulated NOTCH1 signalling at later stages using the *Nfatc1*^{cre} driver, which becomes fully active around E9.0, when touchdown formation is finalized (Extended Data Fig. 9a–l). We first analysed the morphology of *Notch1* loss-of-function (LOF) and GOF mouse embryos generated with *Nfatc1*^{cre}. In *Nfatc1*^{cre}*Notch1*$^{fl/fl}$ embryos, early phases of trabeculation occurred normally but, at 25–28 pairs of somites (ps), basal-to-apical ECM degradation was defective and the ECM bubble encompassed the whole trabecula (Fig. 3m and Extended Data Fig. 9m). Similar to *Tie2*^{cre}*Notch1*$^{fl/fl}$ embryos, trabecular myocardium area was increased from E9.5 to E10.5, after which it decreased (Extended Data Fig. 9n). In *Nfatc1*^{cre}*Notch1*^{GOF} mice, trabeculation was grossly normal, but there was a marked reduction in ECM from E10.5 and premature trabecular termination at E11.5 instead of E14.5 (Fig. 3p and Extended Data Fig. 9o, p). We next investigated the expression of ventricular chamber genes. Expression levels of chamber markers were unchanged in both mutant mice at the phenotypic onset stage (Extended Data Fig. 10a–g, o). However, in *Nfatc1*^{cre}*Notch1*$^{fl/fl}$ embryos, there was downregulation of ECM degradation markers *Adamts1* and neoversican in the endocardium, as well as ectopic expression of ECM synthesis markers *Has2*, *Vcan* and CD44 in the myocardium and upregulation of *Nrg1* and the VEGFA pathway (Fig. 3n, o and Extended Data Fig. 10h–o). In *Nfatc1*^{cre}*Notch1*^{GOF} mice, ECM degradation markers *Adamts1* and neoversican were upregulated in the endocardium (Fig. 3r and Extended Data Fig. 10j, o). Similar to *Tie2*^{cre}*Notch1*^{GOF} embryos, endocardial *Nrg1* downregulation (at E11.5) led to downregulation of *Has2* and CD44 in the myocardium, as well as downregulation of the VEGFA pathway, promoting premature ECM removal (Fig. 3q and Extended Data Fig. 10h–o). Lastly, we assessed proliferation in *Notch1*^{LOF} and *Notch1*^{GOF} embryos generated with *Nfatc1*^{cre}-expressing embryos. Proliferation and trabecular allocation were unchanged (Extended Data Fig. 10p, q). These results confirm ongoing reciprocal interactions between NRG1 and Notch signalling after touchdown formation, and Notch control of basal-to-apical ECM degradation.

Mice with mutations in rearranged L-myc fusion (*Rlf*)[17] show NCC, which is associated with loss of Notch pathway activity in late fetal life[5]. As in Notch genetic mutants, we found persistence of ECM and excessive trabecular myocardium in both ventricles at E14.5[17] (Fig. 3s and Extended Data Fig. 10r, s). Neoversican was downregulated in the endocardium, indicating defective ECM degradation (Fig. 3u and Extended Data Fig. 10t). Expression of the ECM synthesis gene *Has2* in the myocardium was maintained along with NRG1 signalling in the myocardium (Fig. 3t and Extended Data Fig. 10u, v). Therefore, we hypothesize that the collapse of Notch signalling in *Rlf*$^{D28/D28}$ mutant mice alters ECM dynamics and delays trabecular termination, which may contribute to the NCC phenotype (Fig. 4c, d, g, h).

We propose a new model for cardiac trabeculation in mice in which trabeculation starts before heart tube assembly (E8.0), as opposed to at E9.0 as has been suggested previously[2]. Further studies at earlier stages will determine whether trabeculation in mice proceeds through cardiomyocyte extrusion[6]. Notch-dependent ECM degradation is essential for touchdown/dome formation during early phases and for basal-to-apical ECM excavation during trabecular extension and termination[3] (Fig. 4a–d). The NRG1 and Notch pathways work dynamically and reciprocally to control ECM dynamics and trabecular growth, which establish the blueprint for trabecular architecture. Cellular behaviours associated with endocardial sprouting and touchdown formation, and the involvement of VEGFA and Notch pathways, seem reminiscent of angiogenic sprouting in developing vascular beds[14]. Further studies will reveal whether endocardial sprouting is controlled by tip cell/stalk cell fate decisions. Premature degradation of the ECM blocked trabecular growth (Fig. 4i, j), whereas continued ECM synthesis in the absence of degradation specifically in the context of *Notch1* LOF allowed enhanced trabecular growth by increasing cardiomyocyte allocation to the trabecular layer (Fig. 4e–h). The previously described promotion of cardiomyocyte proliferation by both NOTCH1 and NRG pathways[2,4,13] may be contingent on the correct ECM environment, which is itself controlled by these pathways. Further insights may have relevance for cardiac regeneration. Our study reconciles the morphological defects associated with Notch LOF at early[3] (Fig. 4e) and late[5,8,17] (Fig. 4f, g) stages as excessive trabeculation. Results indicate fundamentally different epistatic relationships between relevant signalling pathways than have been described previously. Whereas *Nrg1*, *Bmp10* and *Hand2* have been reported to be Notch-dependent[2,13], these and other chamber genes were downregulated in *Tie2*^{cre}*Notch1*$^{fl/fl}$ mutants only close to lethality (E9.5), and were unaffected in *NFATc1*^{cre}*Notch1*$^{fl/fl}$ mutants at its phenotypic onset stage (also E9.5), suggesting that late changes may involve secondary effects. Our model provides a template for understanding the key morphological and signalling steps in heart development and how these are affected in congenital cardiomyopathies, with potential implications for other tissues and organs controlled by these pathways, as well as human disease classification and the pursuit of novel therapies.

Online content

Any Methods, including any statements of data availability and Nature Research reporting summaries, along with any additional references and Source Data files, are available in the online version of the paper at https://doi.org/10.1038/s41586-018-0110-6.

Received: 17 December 2016; Accepted: 26 March 2018;
Published online 9 May 2018.

1. Moorman, A. F. & Christoffels, V. M. Cardiac chamber formation: development, genes, and evolution. *Physiol. Rev.* **83**, 1223–1267 (2003).
2. Grego-Bessa, J. et al. Notch signaling is essential for ventricular chamber development. *Dev. Cell* **12**, 415–429 (2007).
3. Stankunas, K. et al. Endocardial Brg1 represses *ADAMTS1* to maintain the microenvironment for myocardial morphogenesis. *Dev. Cell* **14**, 298–311 (2008).

4. Lai, D. et al. Neuregulin 1 sustains the gene regulatory network in both trabecular and nontrabecular myocardium. *Circ. Res.* **107**, 715–727 (2010).
5. Luxán, G. et al. Mutations in the NOTCH pathway regulator *MIB1* cause left ventricular noncompaction cardiomyopathy. *Nat. Med.* **19**, 193–201 (2013).
6. Staudt, D. W. et al. High-resolution imaging of cardiomyocyte behavior reveals two distinct steps in ventricular trabeculation. *Development* **141**, 585–593 (2014).
7. Icardo, J. M. & Fernandez-Terán, A. Morphologic study of ventricular trabeculation in the embryonic chick heart. *Acta Anat.* **130**, 264–274 (1987).
8. D'Amato, G. et al. Sequential Notch activation regulates ventricular chamber development. *Nat. Cell Biol.* **18**, 7–20 (2016).
9. Camenisch, T. D. et al. Disruption of hyaluronan synthase-2 abrogates normal cardiac morphogenesis and hyaluronan-mediated transformation of epithelium to mesenchyme. *J. Clin. Invest.* **106**, 349–360 (2000).
10. Hatano, S. et al. Versican/PG-M is essential for ventricular septal formation subsequent to cardiac atrioventricular cushion development. *Glycobiology* **22**, 1268–1277 (2012).
11. Zhou, Z. et al. The cerebral cavernous malformation pathway controls cardiac development via regulation of endocardial MEKK3 signaling and KLF expression. *Dev. Cell* **32**, 168–180 (2015).
12. Gessler, M. et al. Mouse *gridlock*: no aortic coarctation or deficiency, but fatal cardiac defects in *Hey2*$^{-/-}$ mice. *Curr. Biol.* **12**, 1601–1604 (2002).
13. VanDusen, N. J. et al. Hand2 is an essential regulator for two Notch-dependent functions within the embryonic endocardium. *Cell Rep.* **9**, 2071–2083 (2014).
14. Ribatti, D. & Crivellato, E. "Sprouting angiogenesis", a reappraisal. *Dev. Biol.* **372**, 157–165 (2012).
15. Iivanainen, E. et al. Intra- and extracellular signaling by endothelial neuregulin-1. *Exp. Cell Res.* **313**, 2896–2909 (2007).
16. Ferrara, N. et al. Heterozygous embryonic lethality induced by targeted inactivation of the VEGF gene. *Nature* **380**, 439–442 (1996).
17. Bourke, L. M. et al. Loss of rearranged L-Myc fusion (RLF) results in defects in heart development in the mouse. *Differentiation* **94**, 8–20 (2017).

Acknowledgements We thank D. Y. Stainier, J. S. Rasouli, A. V. Cherian, J. M. Polo, F. J. Rossello, G. Chapman, V. Sardesai, B. Shewale, J. O'Rourke, S. Tyler, L. Madigan, A. Ahmad, C. Onie and M. Tondl for their contributions, and J. Moreau and E. Forte for critical review of the manuscript. This work was funded by ARC (DP160104858, DP140101067), NHMRC (APP1118576; 1074386; 573732; 573705; 1032851, CDF 1049980), Foundation Leducq (15 CVD 03) and Perpetual Trust (FR2012/0435). J.L.d.l.P. received grants SAF2016-78370-R and CB16/11/00399 (CIBER CV) from MEIC.

Author contributions G.d.M.-N. was involved in the conceptualization, methodology, investigation, validation, analysis, writing, project administration, visualization and supervision of this project; M.R. carried out bioinformatics analyses; A.A.S.A. provided experimental support; B.W., B.Z., A.A., G.D., E.T., L.M.B., S.K.H., M.E.P., R.H.A., H.C., W.S. and J.L.d.l.P. provided resources. R.P.H. provided supervision, acquired funding and was involved in writing of the manuscript.

Competing interests The authors declare no competing interests.

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0110-6.

Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0110-6.

Reprints and permissions information is available at http://www.nature.com/reprints.

Correspondence and requests for materials should be addressed to G.d. or R.P.H.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

Mouse lines and ethical statement. All experimental procedures were in compliance with international and Australian guidelines. Mouse husbandry, experimental procedures and protocols involving mice were approved by the St Vincent's Hospital/Garvan Institute Animal Ethics Committee. $Nrg1^{tm/tm}$ mice (tm refers to transmembrane domain deletion)[4], $ErbB2^{-/-}$ mice[18], $Rlf^{-/-}$($Rlf^{D28/D28}$) mice[19], $Efnb2^{gfp/gfp}$ mice[20], $BMP10^{-/-}$ mice[21], $Tie2^{cre}$ mice[22], $Tie1^{cre}$ mice[23], $Nfatc1^{cre}$ mice[24], $Notch1^{fl/fl}$ mice[25], $Rbpj^{fl/fl}$ mice[26], $VEGFR2^{fl/fl}$ mice[27], $Notch1^{GOF}$ mice[28], *R26R-LacZ* mice[29], $R26^{mTmG}$ mice[30] were used.

Embryo stage analysis. In this study, embryonic stage was given as pairs of somites (ps). Quantifications of changes in cellular behaviours and morphological areas required analysis of embryos from E8.0 to E8.5 every 2 ps, from E8.5 to E9.0 every 3 ps and from E9.5 to E10.5 every 5 ps. From E11.5, we followed conventional staging. The phenotypic onset stage for molecular analysis for the different mutant lines used in the study was determined by quantitative time-course morphological analysis from E8.0 until near the point of lethality. The phenotypic onset stage was defined as the earliest stage when both the trabecular myocardium and ECM areas were significantly different by quantification. The phenotypic onset stages for molecular analysis were: E9.0 (18–20 ps) for $Nrg1^{tm/tm}$ and $Erbb2^{-/-}$ embryos; E8.5 (13–15 ps) for $Tie2^{cre}Notch1^{fl/fl}$, $Tie2^{cre}Notch1^{GOF}$, $Tie2^{cre}Rbpj^{fl/fl}$ and $Tie1^{cre}Vegfr2^{fl/fl}$ embryos; E10.0 (25–28 ps) for $Nfatc1^{cre}Notch1^{fl/fl}$ embryos; E10.5 (30–33 ps) for $Nfatc1^{cre}Notch1^{GOF}$ embryos and E13.5 for $Rlf^{D28/D28}$ embryos.

Tissue processing. For immunofluorescence, embryos from E7.5–E10.5 were fixed in 4% paraformaldehyde for 2 h at 4 °C. Embryos from E11.5 onwards and all used for in situ hybridization (ISH) and RNA-scope were fixed overnight (12 h) at 4 °C. After fixation, embryos were embedded in paraffin using standard protocols.

Histology, Alcian blue and LacZ staining. Haematoxylin and eosin and Alcian blue stainings were performed following standard protocols. Whole-mount LacZ stainings were performed following a previously described protocol[31].

ISH and RNA-scope. ISH was performed as previously described[32] (details of the used probes are available on request). *Nrg1* ISH using RNA-scope2.5 was performed following the manufacturer's protocol (Advanced Cell Diagnostics) with methyl green counterstaining. Imaging was performed using a Leica DM4000 bright-field microscope.

Immunofluorescence. Immunofluorescence was performed as previously described[33,34] using the following antibodies: BRG1 (sc-17796, Santa Cruz Biotechnology), N1ICD (4147, Cell Signaling Technology), DLL4 (sc-28915, Santa Cruz Biotechnology), FLK1 (sc-48161, Santa Cruz Biotechnology), pAKT Thr308 (2965, Cell Signaling Technology), fibronectin (F3648, Sigma-Aldrich), laminin (L9393, Sigma-Aldrich), collagen IV (ab19808, Abcam) and Ki-67 (ab15580, Abcam). Immunofluorescence for CD44 (559250, BD Pharmigen), neoversican (PA1-1748A, Thermo Scientific), pERBB2 Thr877 (ab47262, Abcam), aggrecan (ab3778, Abcam), perlecan (MAB1948P, Millipore), and collagen I (ab21286, Abcam) was performed similarly but without antigen retrieval. Biotinylated-HABP (385911, Millipore) staining was performed following a similar protocol but without antigen retrieval and the secondary antibody step. The specificity of the pERBB2 antibody was proven by testing in $ErbB2^{-/-}$ and $Nrg1^{tm/tm}$ embryos (Extended Data Fig. 4e, f). All sections were counterstained with the nuclear marker DAPI (422801, Sigma-Aldrich) and muscle marker smooth muscle actin (F3777, SMA, Sigma-Aldrich). Imaging was performed using a Zeiss ImagerZ1/LSM700 microscope.

Whole-mount immunofluorescence and imaging. Whole embryos from wild-type, $Tie2^{cre}Notch1^{fl/fl}$ and $Nrg1^{tm/tm}$ combined with the $Tie2^{cre}R26^{mTmG}$ alleles were stained in whole-mount with anti-GFP antibody (ab13970, Abcam) to enhance the GFP signal. Embryos where permeabilized with 0.3% Triton X-100 in PBS five times for 1 h at room temperature followed by a 1-h incubation in blocking buffer (10% goat serum in 0.3% Triton X-100 in PBS). Overnight incubation was performed with anti-GFP antibody in blocking buffer followed by three washes for 1 h in 0.3% Triton X-100 in PBS. Secondary antibody amplification together with nuclear and myocardium counterstainings were performed overnight by incubation of samples with a mix of anti-chicken secondary antibody coupled to AlexaFluor488 (A11039, ThermoFisher Scientific), nuclear marker DAPI (422801, Sigma-Aldrich) and muscle marker smooth muscle actin (C6198, SMA–Cy3, Sigma-Aldrich). Then, samples were washed for 3 h in 0.3% Triton X-100 in PBS. Embryos were dehydrated in a methanol series (50%, 70%, 90% and 100%) and tissue clearing was performed by incubation of samples in BABB (benzyl alcohol and benzyl benzoate, Sigma-Aldrich). Whole-heart *z*-stacks were captured in a Zeiss Line Scanning Microscope 7DUO.

3D reconstructions and 3D PDFs. Whole-heart confocal *z*-stacks from wild-type, $Tie2^{cre}Notch1^{fl/fl}$ and $Nrg1^{tm/tm}$ embryos combined with the $Tie2^{cre}R26^{mTmG}$ alleles were reconstructed using the Amira 6 software (FEI Visualization Services Group) as previously described[35]. Segmentation of endocardium, myocardium and ECM based on fluorescence signal from the confocal *z*-stack was performed. 3D interactive PDFs were generated using the Adobe Acrobat Professional 9 Extended Version and edited to incorporate interactive tools as previously described[36]. To activate the 3D PDF, any of the default views on the left side can be clicked and the 3D PDF can be navigated using the default views or by manually removing the different tissue layers with the buttons above the models and rotating the model.

Transmission electron microscopy. Embryos used for transmission electron microscopy were fixed and processed following standard protocols. Embryos were fixed in 2.5% glutaraldehyde and 2% paraformaldehyde in 0.1 M phosphate buffer (pH 7.4) overnight and post-fixed in 1% osmium tetroxide in 0.1 M phosphate buffer. After fixation, samples were washed, dehydrated and infiltrated with Epon-Araldite. Ultrathin sections were imaged on a JEOL JEM 1011 transmission electron microscope with Gatan ES1000W camera.

RBPj-binding site identification in *Adamts1* and *Mmp2* promoter and enhancer regions. In silico bioinformatics analysis was performed using the UCSC Genome Browser (http://www.genome.ucsc.edu/)[37] on genomic regions surrounding the *Adamts1* locus to identify potential regulatory regions containing RBPj-binding sites. Histone modifications patterns within intergenic regions between genes of interest and immediately neighbouring genes were used to predict active enhancers (H3K4me1/H3K27ac) and promoters (H3K4me3/H3K27ac) in endothelial or cardiac tissue[38]. Putative regulatory regions bound by cardiac kernel transcription factors, such as NKX2-5, GATA4 or p300[38], were further prioritized. This preliminary analysis allowed us to identify the promoter (mm10_dna range = chr16:85803133-85806294) and a potential enhancer region (mm10_dna range = chr16:85754728-85757691) downstream of the *Adamts1* gene. RBPj motif analysis was then applied to these regions using Possum (https://www.zlab.bu.edu/~mfrith/possum). All binding sites that were found were filtered based on similarity scores to the consensus RBPj-binding motif matrix (http://ccg.vital-it.ch/cgi-bin/pwmtools/pwmviewer.cgi?ID=RBPJ.p2&lib=swissregulon) and on sequence evolutionary conservation among vertebrates. The analysis reduced the candidates to three RBPj-binding sites in the *Adamts1* promoter region and one in a putative *Adamts1* downstream enhancer region.

N1ICD chromatin immunoprecipitation in mouse aortic endothelial cells. Chromatin immunoprecipitation was performed on mouse aortic endothelial cells that were maintained as described[39]. No authentication was performed. This line tested negative for mycoplasma contamination. Approximately 7 million cells were trypsinized and pelleted, resuspended in 1% formaldehyde in PBS and fixed at room temperature for 10 min with gentle rotation followed by two washes in PBS with final resuspension in lysis buffer (50 mM HEPES-KOH pH 7.5, 140 mM NaCl, 1 mM EDTA pH 8, 0.1% SDS, 1% Triton X-100, 0.1% sodium deoxycholate and protease inhibitors). Sonication was performed in a Bioruptor Pico sonicator (Diagenode) for six cycles of 10 repeats (30 s on, 30 s off). The DNA input sample to determine DNA concentration was treated with RNaseA (0.5 mg/ml) at 37 °C for 30 min followed by proteinase K ($20\,\mathrm{mg\,ml^{-1}}$) treatment at 65 °C for 2 h and final purification using QIAquick PCR purification kit (28106, Qiagen). Sonicated samples were divided into N1ICD, IgG, H3K4 and beads-only conditions containing around 6 μg of DNA per condition. Immunoprecipitation was performed by diluting samples in RIPA buffer (50 mM Tris-HCl pH 8, 150 mM NaCl, 2 mM EDTA, 1% IGEPAL CA-630, 0.1% SDS, 0.5% sodium deoxycholate and protease inhibitors). Antibodies (N1ICD (1:200, 4147 Cell Signaling Technology), IgG control (5 μg, sc-2027 Santa Cruz Biotechnology), H3K4 (ab3778, Abcam) and sepharose beads (17-0618-01, Sigma-Aldrich) were added and incubated overnight with rocking at 4 °C. After antibody incubation, samples were washed five times for 15 min in wash buffer (0.1% SDS, 2 mM EDTA pH 8, 150 mM NaCl, 20 mM Tris-HCl pH 8, 1% Triton X-100) followed by two washes for 15 min in final wash buffer (0.1% SDS, 2 mM EDTA pH 8, 500 mM NaCl, 20 mM Tris-HCl pH 8, 1% Triton X-100). Release of precipitated DNA was performed with elution buffer (1% SDS, 100 mM $NaHCO_3$) with DNA purification using the QIAquick PCR purification kit (28106, Qiagen). Analysis of DNA sequences binding to N1ICD was performed by PCR using specific primers flanking RBPj-binding sites of interest. Primers: *Adamts1* PR1: F: 5′-TCGTCACTCTGGGGGTGTAA-3′, R: 5′-TGGTGCCCATGGAAACGTAG-3′; *Adamts1* PR2: F: 5′-TCCTTTGCTCCTCCCCAGAT-3′, R: 5′-AGGTTGGAGGTTGGGGTAGT-3′; *Adamts1* PR3: F: 5′-AAAGGGGAAAGGGTTGCCAG-3′, R: 5′-TGTGCCAAACACACAATGCT-3′; *Adamts1* DSR: F: 5′-AAACCATTGGATCGAGGCCC-3′, R: 5′-TGACGCTAAGGAGAACGCAA-3′; positive control *Efnb2 bsA* (the binding site in the promoter region of *Efnb2*) F: 5′-AACAGCGCATGGAAACTACC-3′, R: 5′-CCTTCCTGGGTCTCTTAGGC-3′[2] PCR conditions: 94 °C for 1 min, 35 rounds of 98 °C for 10 s, 60 °C for 30 s, 72 °C for 30 s, followed by 72 °C for 3 min.

Cloning, mutagenesis and analysis of *Adamts1* promoter and enhancer regions. The promoter of *Adamts1* was PCR-amplified and cloned into a pGL4.10-luc2 vector (Promega). The *Adamts1* downstream enhancer region was cloned into a pGL4.23-luc2/minP vector (Promega). Primers used: *Adamts1* promoter: F: 5′-ATATAAGAGCTCTGAGCTCAGTCGGTGCTAAAC-3′, R: 5′-ATATAACTCGAGAAGACGCGTTAAGCTGGTGT-3′; *Adamts1* downstream enhancer: F: 5′-ATATAAGAGCTCCAGGGACTGGCATTAAAAGTTCT-3′,

R: 5′-ATATAACTCGAGCTG TGATTCTGGGCTTCTAAGGT-3′. PCR mutagenesis of RBPj-binding sites within promoter/enhancer regions used the following primers: *Adamts1* promoter binding site 1: F: 5′-CGAGAGATGAAGTTAAAGGTTCACTCTTTGATAACAAGCAGCCAGCGCTGTACAAAGTGCTGGTC-3′, R: 5′-GACCAGCACTTTGTACAGCGCTGGCTGCTTGTTATCAAAGAGTGAACCTTTAACTTCATCTCTCG-3′; *Adamts1* promoter binding site 2: F: 5′-TCTTCTTTTAATTCGATATTCCCTGGTACTTAGCACAAAGTTTTTGTTTGTTTTCATGAAGACTATCTAGAACCCTGGGAAAGTG-3′, R: 5′-CACTTTCCCAGGGTTCTAGATAGTCTTCATGAAAACAAACAAAAACTTTGTGCTAAGTACCAGGGAATATCGAATTAAAAGAAGA-3′; *Adamts1* promoter binding site 3: F: 5′-CAGATGGGTATGTGTTATAGGCTTGTTCAATGCACACACAATGCTGATCTTAATTGTTTTAAAAGTGGATTAAGACTCTTCTG-3′, R: 5′-CAGAAGAGTCTCTAATCCACTTTTAAAACAATTAAGATCAGCATTGTGTGTGCATTGAACAAGCCTATAACACATACCCATCTG-3′; *Adamts1* downstream enhancer: F: 5′-CCTTGGATTTTTGTTTTCTTTTCCAAGTGCTTAGCAGAAAATAAAACTGACAGTCTCAAGACTGGTAATCTGAGTCAAAGGGG-3′, R: 5′-CCCCTTTGACTCAGATTACCAGTCTTGAGACTGTCAGTTTTATTTTCTGCTAAGCACTTGGAAAAGAAAACAAAAATCCAAGG-3′. Luciferase assays were performed in C2C12 cells (DMEM high glucose (ThermoFisher) supplemented with 10% FCS), which have an active Notch pathway. No authentication was performed. This line tested negative for mycoplasma contamination. Transfection was performed using Lipofectamine LTX (ThermoFisher) in Opti-MEM (ThermoFisher), 150 ng of luciferase reporter vector (wild-type or mutated promoter/enhancer), 150 ng of pCMX-N1ICD plasmid or empty vector, and 10 ng of pCMX-Renila as a transfection control. As a control for NOTCH1 activity we used a HEY1-Luc Notch activity reporter[40]. The transfection mix was added to the cells and incubated for 5 h followed by a change of medium. 24 h after the medium change, cells were lysed using buffer from the Dual-Luciferase Reporter Assay System kit (Promega) and luciferase activity was measured using a FLUOstar-OPTIMA luminometer (BMG).

RBEC and rat serum production. Embryo cultures were performed as described[41] with modifications. Embryos were cultured in a RBEC precision incubator (BT Engineering) in DMEM high glucose medium (Life Technologies) and rat serum (50%) with optimized concentrations of inhibitors or cytokines for 24 h. Rat serum was produced by descending aorta puncture using standard protocols. Hyaluronidase treatments: 20 turbidity reducing units (TRU) of hyaluronidase (Sigma-Aldrich) as previously decribed[42]; metalloprotease activity inhibition: 20 μM GM6001 (Milipore) as previously described[3]. NRG1 treatments: 25 nM human recombinant NRG1 (R&D Systems). Notch pathway inhibition: 50 μM DAPT (Sigma-Aldrich) as previously described[2]. NRG1 pathway inhibition: 5 μM PD158780 (Calbiochem)[43] after determining the optimal concentration to induce a $Nrg1^{tm/tm}$ mutant-like phenotype. VEGFA treatments: 2.5 μM VEGFA (R&D Systems). Inhibition of VEGFR2: 10 μM SU5416 (Sigma-Aldrich).

RNA extraction and quantitative PCR. The experimental samples were generated by pooling three dissected left ventricles from $Nfatc1^{cre}Notch1^{fl/fl}$ and $Nfatc1^{cre}Notch1^{GOF}$ embryos and six for $Tie2^{cre}Notch1^{fl/fl}$, $Tie2^{cre}Notch1^{GOF}$ and $Nrg1^{tm/tm}$ embryos. Each pooled sample constituted one biological replicate and a minimum of three biological replicates were analysed for each wild-type or mutant condition. qPCR assays were performed in experimental triplicates on each biological replicate. RNA extraction was performed using TRIzol (Life Technologies) following standard protocols. cDNA was generated using SuperScript-III First-Strand Synthesis System (ThermoFisher Scientific). Quantitative PCR was performed in experimental triplicates by using Sybr Green (Roche) in a Lightcycler480 (Roche). Primers used: *Has2*[3] F: 5′-CGTGGATTATGTACAGGTGTGTGA-3′, R: 5′-CTGAGGAAGGAGATCCAGGAAT-3′; *Vcan*[3] F: 5′-CCTGATAGCAGATTTGATGCCTA-3′, R: 5′-CTAATGGAATGACTGGTGTAGCTC-3′; *Adamts1*[3] F: 5′-CTGGGCAAGAAATCTGATGA-3′, R: 5′-AAGCACAGCCACAGTTTATCA-3′; *Adamts4* (this study) F: 5′-ATGGCCTCAATCCATCCCAG-3′ R: 5′-AAGCAGGGTTGGAATCTTTGC-3′; *Adamts5* (this study) F: 5′-GGAGCGAGGCCATTTACAAC-3′, R: 5′-CGTAGACAAGGTAGCCCACTTT-3′; *Adamts8* (this study) F: 5′-AGGTGGAGATGGGTAATGGG-3′, R: 5′-CACTTTGCGGGAGTCTTCTG-3′; *Adamts9* (this study) F: 5′-GACTTGTGGGCAAGGTAAGG-3′, R: 5′-TCAGTCTCGGGGATGTAATCTG-3′; *Mmp2* (this study) F: 5′-CAAGTTCCCCGGCGATGTC-3′, R: 5′-TTCTGGTCAAGGTCACCTGTC-3′; *Hyal1* (this study) F: 5′-ACCTGCTTCGCATCTCTACTC-3′, R: 5′-GGTTGGATACCACGGAACCTC-3′; *Hyal2* (this study) F: 5′-AGATGTTTACCGACAGTCTTCAC-3′, R: 5′-TGACGTAACGGAGAGTGTTCAA-3′; *Efnb2* (this study) F: 5′-ATTATTTGCCCCAAAGTGGACTC-3′, R: 5′-GCAGCGGGGTATTCTCCTTC-3′; *Bmp10* (this study) F: 5′-ATGGGGTCTCTGGTTCTGC-3′, R: 5′-CAATACCATCTTGCTCCGTGAA-3′; *Notch1* (this study) F: 5′-GATGGCCTCAATGGGTACAAG-3′, R: 5′-TCGTTGTTGTTGATGTCACAGT-3′; *Dll4* (this study) F: 5′-TTCCAGGCAACCTTCTCCGA-3′, R: 5′-ACTGCCGCTATTCTTGTCCC-3′; *Nrg1* (this study) F: 5′-TTCCCATTCTGGCTTGTCTAGT-3′, R: 5′-CCAGGGTCAAGGTGGGTAG-3′; *ErbB2* (this study) F: 5′-GAGACAGAGCTAAGGAAGCTGA-3′, R: 5′-ACGGGGATTTTCACGTTCTCC-3′; *Vegfa* (this study) F: 5′-GCACATAGAGAGAATGAGCTTCC-3′, R: 5′-CTCCGC TCTGAACAAGGCT-3′; *Vegfr2*[3] F: 5′-CCCCAAATTCCATTATGACAAC-3′, R: 5′-CTGGGATCACTTTTACTTCTGGTT-3′; *Hand2* (this study) F: 5′-GCAGGACTCAGAGCATCAACA-3′, R: 5′-AGGTAGGCGATGTATCTGGTG-3′; *Nppa* (also known as *Anf*) (this study) F: 5′-GCTTCCAGGCCATATTGGAG-3′, R: 5′-GGGGGCATGACCTCATCTT-3′; *Mest* (this study) F: 5′-GTGGTGGGTCCAAGTAGGG-3′, R: 5′-AAGCACAACTATCTCAGGGCT-3′; *Hopx* (this study) F: 5′-ACCACGCTGTGCCTCATC-3′, R: 5′-GCGCTGCTTAAACCATTTCT-3′; *Gja5* (this study) F: 5′-GGTCCACAAGCACTCCACAG-3′, R: 5′-CTGAATGGTATCGCACCGGAA-3′; *Hey2* (this study) F: 5′-CGGTGAACAAGCGAGAGTC-3′, R: 5′-GGGGCAGTAGCTGGAACAG–3′; *Mycn* (This study) F: 5′-CGGTGAACAAGCGAGAGTC-3′, R: 5′-GGGGCAGTAGCTGGAACAG-3′; *Hprt*[3] F: 5′-GCTGGTGAAAAGGACCTCT-3′, R: 5′-CACAGGACT AGAACACCTGC-3′; *Gapdh* (this study) F: 5′-ATGGGGAAGGTGAAGGTCG-3′, R: 5′-TAAAAGCAGCCCTGGTGACC-3′; 18S rRNA (this study) F: 5′-CGGCGACGACCCATTCGAAC-3′, R: 5′-GAATCGAACCCTGATTCCCCGTC-3′.

Proliferation and cell quantifications. Proliferation was analysed by immunofluorescence analysis of Ki-67 expression (ab15580, Abcam) followed by segmentation of Ki-67 positive and negative nuclei using Amira 6 software (FEI Visualization Services Group, USA). Trabecular cardiomyocyte proportions were calculated by determining the number of cells included in the trabecular myocardium compared to the total number of cells.

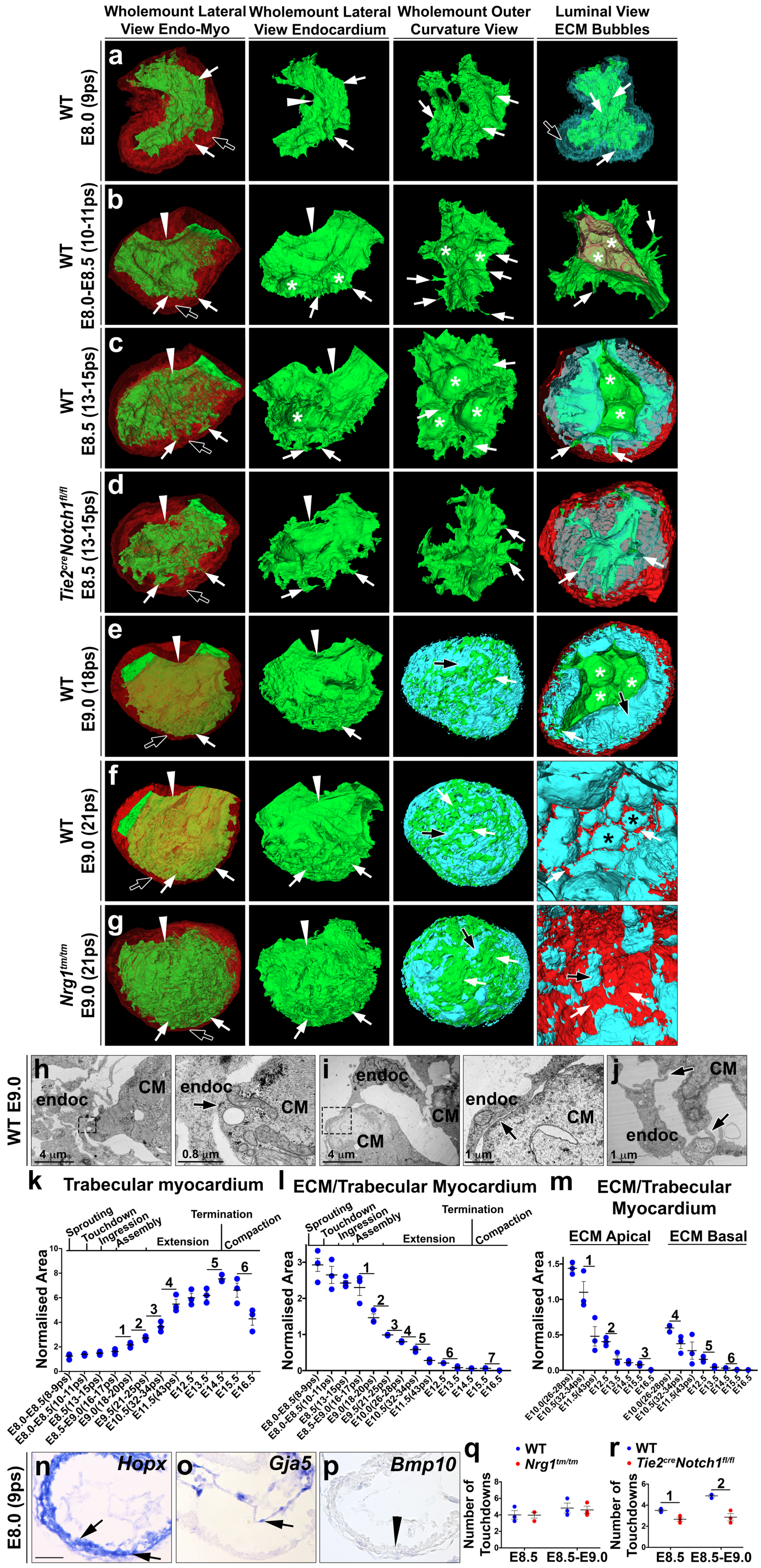
Quantification and statistics. Morphometric quantification was performed using Amira 6 (FEI Visualization Services Group). Two consecutive mid-ventricle H&E-stained sections from embryos at 6–9 ps or 2–4 consecutive images for embryos at 10 ps and later stages were selected for quantification. After light inversion performed in Photoshop (Adobe), images were saved in Amira 6. Segmentation used a combination of automatic and manual selections available in the software. The different cellular tissues (endocardium, trabecular myocardium and compact myocardium) and ECM were separated in different selections for independent area quantification. The criteria applied to segment the different tissues is based on the fact that the compact myocardium is a monolayer until E9.5–E10.0 and therefore the trabecular myocardium consist of all the cardiomyocytes positioned luminal to the compact myocardium monolayer. Furthermore, the endocardial touchdowns always made stable contacts to the compact myocardium indicating that at later stages, when the compact myocardium becomes multilayered, the endocardium at the base of the trabeculae delimitates the separation between the trabecular and the compact myocardium layers. The final area was normalized to the outer circumference of the chamber region analysed. In order to highlight differences between apical and basal trabecular regions, trabecular myocardium with associated ECM was separated into halves along the radial axis. Similar criteria were applied to the mutant embryos analysed. N1ICD and BRG1 quantifications were performed with ImageJ by counting total and N1ICD or BRG1 positive nuclei in the left ventricle from at least three consecutive sections. All quantifications in mouse embryos were performed in at least three different wild-type or mutant embryos. In RBECs, a minimum of five embryos per condition were analysed. The resulting area was normalized to the perimeter of the ventricular area analysed. No statistical methods were used to calculate sample size. Sample size was chosen empirically according to the variability observed during morphological characterization of the different wild-type and mutant phenotypes, and during similar analyses on control and treated embryos in RBEC assays. Exclusion of samples from experimental repeats was only applied in RBEC assays when control wild-type or mutant samples showed unexpected heart defects due to suboptimal culturing conditions. In these cases, the whole experiment was excluded before quantification. No randomization or blinding was applied during the study. Statistical analysis was performed using Prism 6 (Graphpad). After confirming normal distribution by D'Agostino–Pearson omnibus normality test, significance was determined using two-sided Student's *t*-tests for each dataset. Quantitative data are shown as mean ± s.e.m. Variance comparisons were calculated using *F*-tests between the groups that were compared. Variances between statistically compared groups were not always similar. Unpaired Student's *t*-tests with Welch's correction were used to determine the significance of the results when groups showed different variances.

Reporting summary. Further information on experimental design is available in the Nature Research Reporting Summary linked to this paper.

Data availability. The authors declare that the data supporting the findings of this study are available within the paper and its Supplementary Information. Source Data for Figs. 1–3 and Extended Data Figs. 1, 4–10 are available in the online version of the paper.

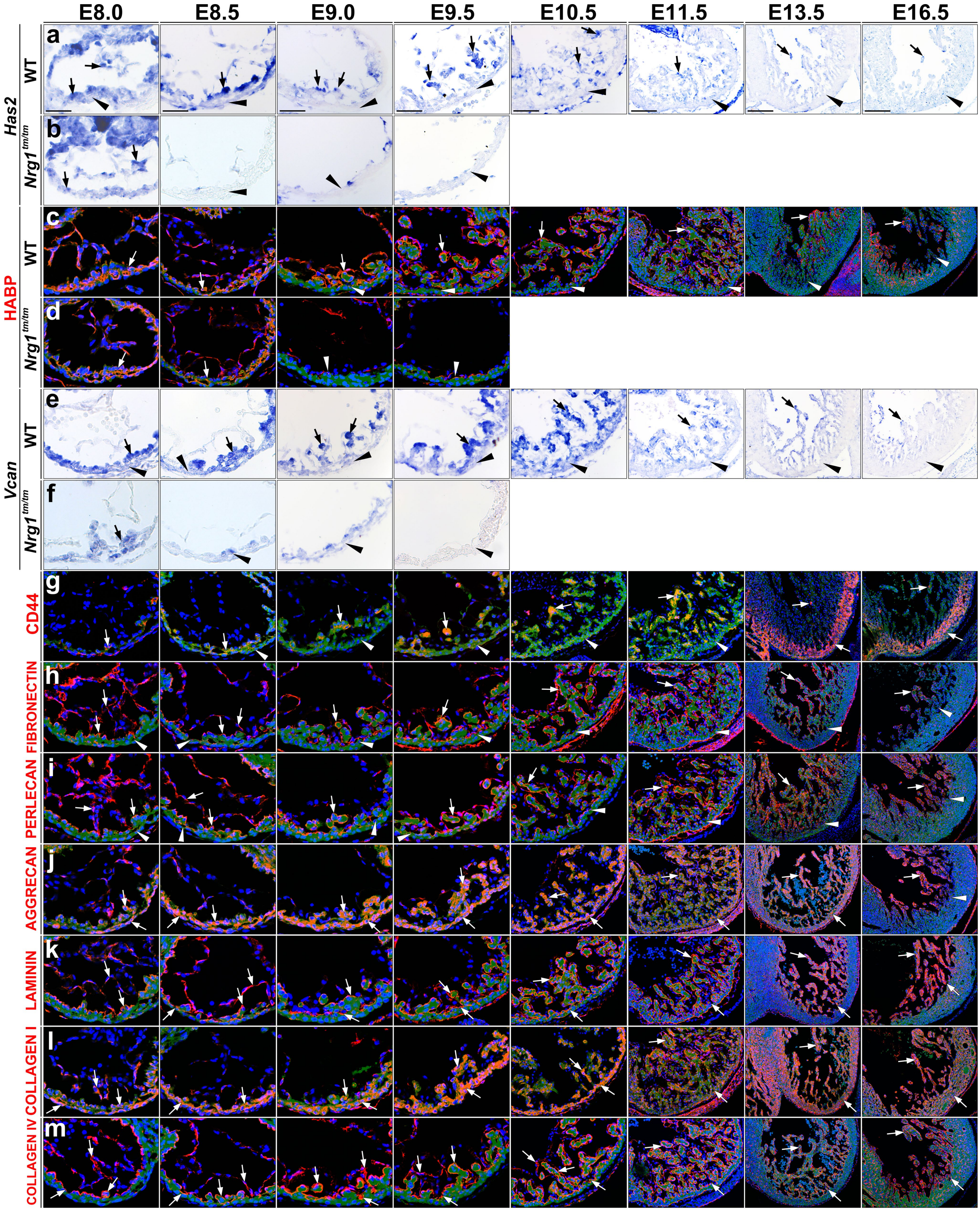
18. Britsch, S. et al. The ErbB2 and ErbB3 receptors and their ligand, neuregulin-1, are essential for development of the sympathetic nervous system. *Genes Dev.* **12**, 1825–1836 (1998).

19. Daxinger, L. et al. An ENU mutagenesis screen identifies novel and known genes involved in epigenetic processes in the mouse. *Genome Biol.* **14**, R96 (2013).
20. Davy, A. & Soriano, P. Ephrin-B2 forward signaling regulates somite patterning and neural crest cell development. *Dev. Biol.* **304**, 182–193 (2007).
21. Chen, H. et al. BMP10 is essential for maintaining cardiac growth during murine cardiogenesis. *Development* **131**, 2219–2231 (2004).
22. Koni, P. A. et al. Conditional vascular cell adhesion molecule 1 deletion in mice: impaired lymphocyte migration to bone marrow. *J. Exp. Med.* **193**, 741–754 (2001).
23. Gustafsson, E., Brakebusch, C., Hietanen, K. & Fässler, R. Tie-1-directed expression of Cre recombinase in endothelial cells of embryoid bodies and transgenic mice. *J. Cell Sci.* **114**, 671–676 (2001).
24. Wu, B. et al. Endocardial cells form the coronary arteries by angiogenesis through myocardial–endocardial VEGF signaling. *Cell* **151**, 1083–1096 (2012).
25. Radtke, F. et al. Deficient T cell fate specification in mice with an induced inactivation of Notch1. *Immunity* **10**, 547–558 (1999).
26. Han, H. et al. Inducible gene knockout of transcription factor recombination signal binding protein-J reveals its essential role in T versus B lineage decision. *Int. Immunol.* **14**, 637–645 (2002).
27. Haigh, J. J. et al. Cortical and retinal defects caused by dosage-dependent reductions in VEGF-A paracrine signaling. *Dev. Biol.* **262**, 225–241 (2003).
28. Watanabe, Y. et al. Activation of Notch1 signaling in cardiogenic mesoderm induces abnormal heart morphogenesis in mouse. *Development* **133**, 1625–1634 (2006).
29. Soriano, P. Generalized *lacZ* expression with the ROSA26 Cre reporter strain. *Nat. Genet.* **21**, 70–71 (1999).
30. Muzumdar, M. D., Tasic, B., Miyamichi, K., Li, L. & Luo, L. A global double-fluorescent Cre reporter mouse. *Genesis* **45**, 593–605 (2007).
31. Shalaby, F. et al. Failure of blood-island formation and vasculogenesis in Flk-1-deficient mice. *Nature* **376**, 62–66 (1995).
32. Kanzler, B., Kuschert, S. J., Liu, Y. H. & Mallo, M. Hoxa-2 restricts the chondrogenic domain and inhibits bone formation during development of the branchial area. *Development* **125**, 2587–2597 (1998).
33. Del Monte, G., Grego-Bessa, J., González-Rajal, A., Bolós, V. & De La Pompa, J. L. Monitoring Notch1 activity in development: evidence for a feedback regulatory loop. *Dev. Dyn.* **236**, 2594–2614 (2007).
34. del Monte, G. et al. Differential Notch signaling in the epicardium is required for cardiac inflow development and coronary vessel morphogenesis. *Circ. Res.* **108**, 824–836 (2011).
35. Soufan, A. T. et al. Three-dimensional reconstruction of gene expression patterns during cardiac development. *Physiol. Genomics* **13**, 187–195 (2003).
36. de Boer, B. A. et al. The interactive presentation of 3D information obtained from reconstructed datasets and 3D placement of single histological sections with the 3D portable document format. *Development* **138**, 159–167 (2011).
37. Kent, W. J. et al. The human genome browser at UCSC. *Genome Res.* **12**, 996–1006 (2002).
38. The ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
39. Nishiyama, T. et al. Functional analysis of an established mouse vascular endothelial cell line. *J. Vasc. Res.* **44**, 138–148 (2007).
40. Maier, M. M. & Gessler, M. Comparative analysis of the human and mouse *Hey1* promoter: *Hey* genes are new Notch target genes. *Biochem. Biophys. Res. Commun.* **275**, 652–660 (2000).
41. Sturm, K. & Tam, P. P. Isolation and culture of whole postimplantation embryos and germ layer derivatives. *Methods Enzymol.* **225**, 164–190 (1993).
42. Baldwin, H. S. & Solursh, M. Degradation of hyaluronic acid does not prevent looping of the mammalian heart in situ. *Dev. Biol.* **136**, 555–559 (1989).
43. Budi, E. H., Patterson, L. B. & Parichy, D. M. Embryonic requirements for ErbB signaling in neural crest development and adult pigment pattern formation. *Development* **135**, 2603–2614 (2008).



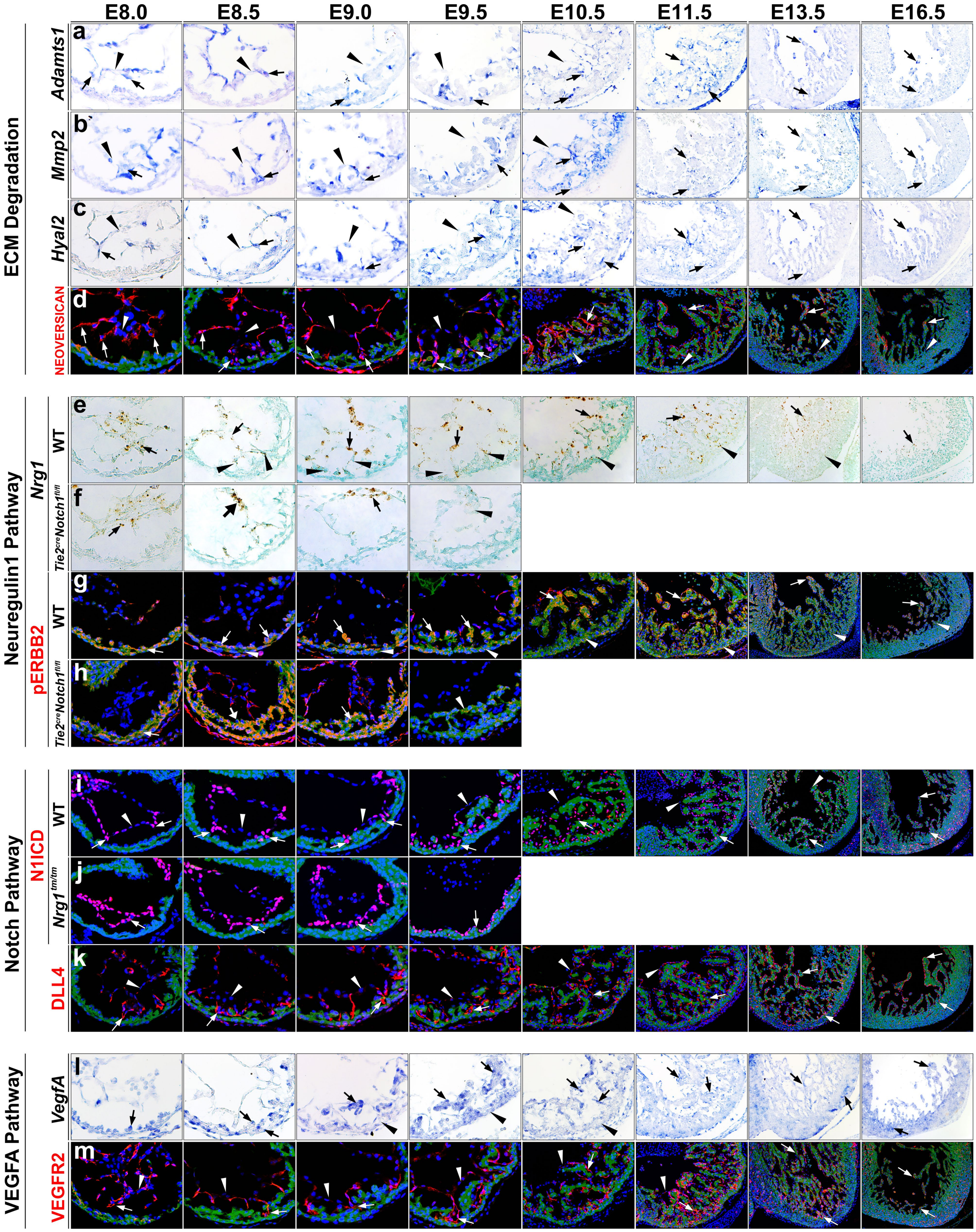
Extended Data Fig. 1 | See next page for caption.

Extended Data Fig. 1 | Analysis of trabeculation by 3D morphological time course reconstructions and electron microscopy. Related to Fig. 1. **a**–**g**, Time-course of left ventricle 3D reconstructions (see Supplementary Figs. 1–7) of wild-type mice (**a**–**c**, **e**, **f**), *Tie2*^{cre}*Notch1*$^{fl/fl}$ mice at E8.5 (**d**) and *Nrg1*$^{tm/tm}$ mice at E9.0 (**g**). In all panels: endocardium (green), myocardium (red, transparent red in a–g (left)), ECM (blue, transparent blue in the right panels of **a**, **c**, **d**). Lateral view of the whole heart (**a**–**g**, left) or endocardium only (**a**–**g**, second left) showing that the endocardial behaviours described in the study occur exclusively in the ventricular outer curvature (white arrows) and not in the inner curvature (arrowheads), and the progressive ECM reduction seen as the distance between the endocardium and the ventricular outer layer as trabeculation proceeds (black arrows). This distance is greater in *Tie2*^{cre}*Notch1*$^{fl/fl}$ (**d**) and smaller in *Nrg1*$^{tm/tm}$ (**g**) hearts. **a**–**g**, Second from the right, view towards the endocardium from the outer curvature. **a**–**d**, Second from the right, note the formation of endocardial ridges between the leading touchdowns (white arrows) generating endocardial domes (white asterisks). **d**, Second from the right, incomplete touchdown formation and lack of endocardial domes in *Tie2*^{cre}*Notch1*$^{fl/fl}$ hearts. **e**–**g**, Second from the right, outer curvature view showing the contact points between endocardium and myocardium (green, white arrow) without ECM, and ECM bubbles (blue, black arrow) in wild-type (**e**, **f**), and the increased endocardial contacts and decreased ECM in *Nrg1*$^{tm/tm}$ (**g**) hearts. **a**–**g**, Right, luminal view of whole heart (**a**) or bisected hearts (**b**–**g**) showing ventricular chamber segmentation by endocardial domes/ECM bubbles (asterisks) in wild-type hearts (**a**–**c**, **e**, **f**) and its defects in *Tie2*^{cre}*Notch1*$^{fl/fl}$ (**d**) and *Nrg1*$^{tm/tm}$ (**g**) hearts. **h**–**j**, Electron microscopy images of wild-type ventricles at E9.0 showing endocardium (endoc) and compact myocardium interaction at the trabecular base. **h**, Desmosomes (arrow). **i**, Endocardial cell in close apposition to a cardiomyocyte (CM). **j**, Cellular projections originating from both cell types. **k**–**m**, Area quantifications on wild-type left ventricle showing trabecular myocardium area normalized to chamber perimeter (**k**), total ECM area (**l**) and ECM area of apical and basal trabecular regions (**m**) normalized to trabecular myocardium. **n**–**p**, ISH of *Hopx* (**n**), *Gja5* (**o**) and *Bmp10* (**p**) in heart sections of wild-type embryos at E8.0. Marker expression (arrow), no marker expression (arrowhead). **q**, **r**, Quantification of touchdown numbers in *Nrg1*$^{tm/tm}$ (**q**) and *Tie2*^{cre}*Notch1*$^{fl/fl}$ (**r**) compared to wild-type embryos at E8.5 and E8.5–E9.0. **a**–**r**, $n = 3$ independent embryos per genotype, except for **k**, **l**, $n = 4$ E8.0 (8–9 ps) and E8.5 (13–15 ps). Quantitative data are shown as mean ± s.e.m. Two-sided Student's t-tests (without corrections for multiple comparisons) were used. Significant comparisons are indicated by numbers as follows. **k**, 1, $P = 0.035$; 2, $P = 0.025$; 3, $P = 0.046$ 4, $P = 0.023$; 5, $P = 0.037$; 6, $P = 0.041$. **l**, 1, $P = 0.046$; 2, $P = 0.046$; 3, $P = 0.0003$; 4, $P = 0.013$; 5, $P = 0.0025$; 6, $P = 0.026$; 7, $P = 0.005$. **m**, 1, $P = 0.038$; 2, $P = 0.013$; 3, $P = 0.0048$; 4, $P = 0.041$; 5, $P = 0.012$; 6, $P = 0.0024$. **r**, 1, $P = 0.018$; 2, $P = 0.0048$. Scale bars, as indicated (h–j) and 20 μm in **n**–**p**.



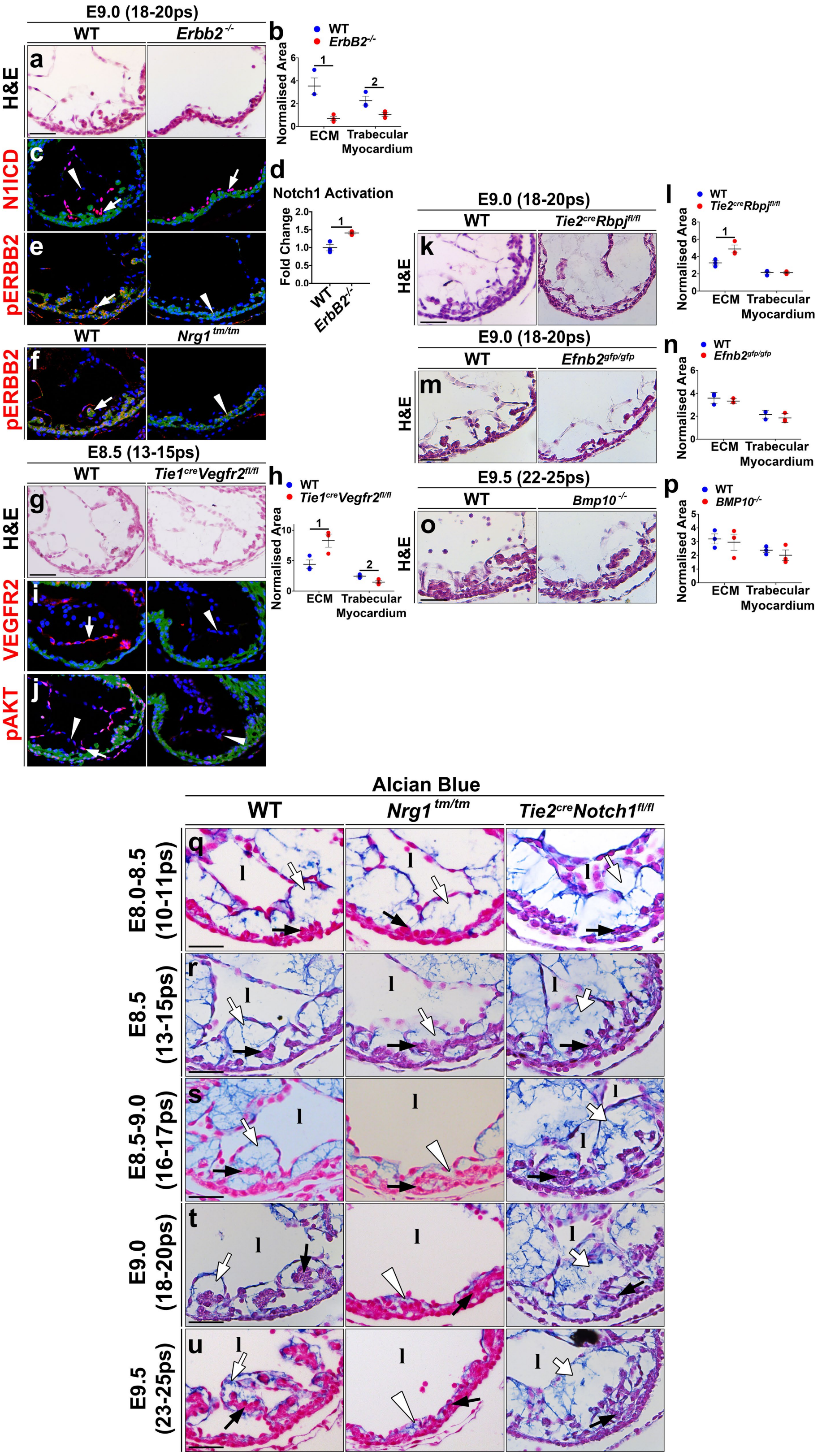
Extended Data Fig. 2 | See next page for caption.

Extended Data Fig. 2 | Time course expression analysis of ECM synthesis markers. Related to Figs. 1, 2. **a**–**m**, Detailed time-course analysis of marker expression using ISH and immunofluorescence in embryonic left ventricle at E8.0, E8.5, E9.0, E9.5, E10.5, E11.5, E13.5 and E16.5 in wild-type embryos (**a**, **c**, **e**, **g**–**m**) as well as in *Nrg1*$^{tm/tm}$ mutants (**b**, **d**, **f**) relative to somite-matched wild-type embryos at E8.0, E8.5, E9.0 and E9.5. **a**, **b**, **e**, **f**, Analysis of *Has2* (**a**, **b**) and *Vcan* (**e**, **f**) by ISH in wild-type (**a**, **e**) and *Nrg1*$^{tm/tm}$ embryos (**b**, **f**). **c**, **d**, Immunofluorescence analysis of HABP in wild-type (**c**) and *Nrg1*$^{tm/tm}$ (**d**) embryos. **g**–**m**, Immunofluorescence analysis of CD44 (**g**), fibronectin (**h**), perlecan (**i**), aggrecan (**j**), laminin (**k**), collagen type I (**l**) and collagen type IV (**m**) in wild-type embryos. Perlecan showed a similar pattern of enrichment in trabecular myocardium and endocardium as HABP and fibronectin, although expression persisted after termination. Aggrecan was expressed in both compact and trabecular myocardium and only became restricted to trabeculae after termination. Laminin and collagen types I and IV did not show restriction. In all panels showing gene/protein expression: marker expression (arrow), reduced marker expression (arrowhead), increased marker expression (thick arrow). In all immunofluorescence images: markers associated with endocardium or ECM (red); markers associated with myocardium (orange); myocardium is stained for SMA (green dulled mask); nuclei (DAPI, blue). **a**–**m**, $n = 3$ independent embryos were used per genotype. Scale bars, 20 μm (E8.0), 25 μm (E8.5), 30 μm (E9.0, E9.5), 50 μm (E10.5), 60 μm (E11.5), 120 μm (E13.5) and 200 μm (E16.5).



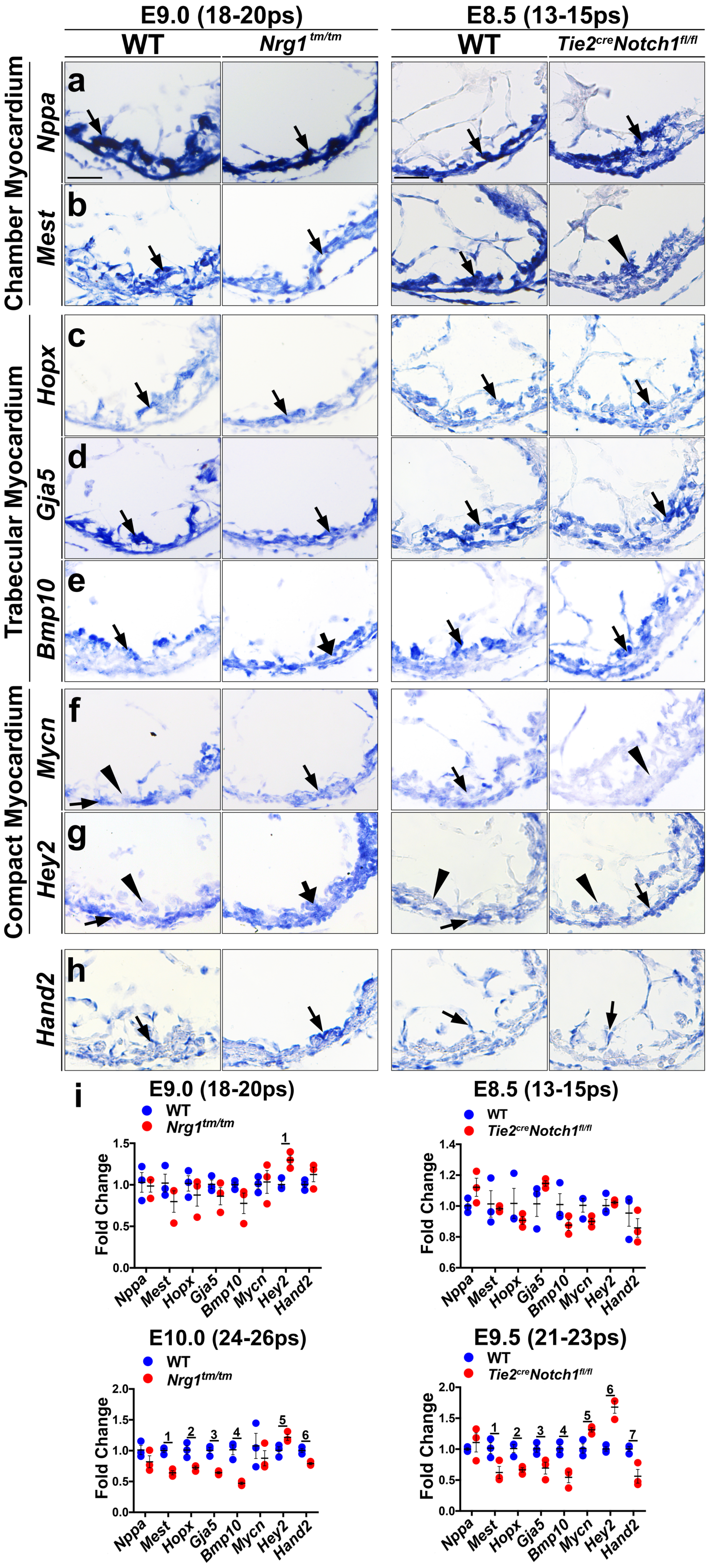
Extended Data Fig. 3 | See next page for caption.

Extended Data Fig. 3 | Time course expression analysis of ECM degradation, NOTCH, NRG1 and VEGFA pathway markers. Related to Figs. 1–3. **a**–**m**, Detailed time-course analysis of marker expression using ISH and immunofluorescence in embryonic left ventricle at E8.0, E8.5, E9.0, E9.5, E10.5, E11.5, E13.5 and E16.5 in wild-type embryos (**a**–**e**, **g**, **i**, **k–m**) as well as in *Tie2*^{cre}*Notch1*$^{fl/fl}$ (**f**, **h**) and *Nrg1*$^{tm/tm}$ (**j**) embryos relative to somite-matched wild-type embryos at E8.0, E8.5, E9.0 and E9.5. Data are organized according to respective ECM degradation, Notch, neuregulin-1 (NRG1) and VEGFA pathways. **a**–**d**, Analysis of ECM degradation markers *Adamts1* (**a**), *Mmp2* (**b**) and *Hyal2* (**c**) by ISH and neoversican (**d**) by immunofluorescence. **e**–**h**, NRG1 pathway analysis of *Nrg1* by RNA-scope (**e**, **f**) and pERBB2 (**g**, **h**) by immunofluorescence in wild-type (**e**, **g**) and *Tie2*^{cre}*Notch1*$^{fl/fl}$ (**f**, **h**) embryos. **i**–**k**, Notch pathway analysis of N1ICD (**i**, **j**) in wild-type (**i**) and *Nrg1*$^{tm/tm}$ (**j**) embryos, and DLL4 in wild-type embryos (**k**) by immunofluorescence. **l**, **m**, VEGFA pathway analysis of *Vegfa* by ISH (**l**) and VEGFR2 by immunofluorescence (**m**). In all panels showing gene/protein expression: marker expression (arrow), reduced marker expression (arrowhead), increased marker expression (thick arrow). In all immunofluorescence images: markers associated with endocardium or ECM (red); markers associated with myocardium (orange); myocardium is stained for SMA (green dulled mask); nuclei (DAPI, blue). **a**–**m**, $n = 3$ independent embryos were used for each genotype. Scale bars, 20 μm (E8.0), 25 μm (E8.5), 30 μm (E9.0, E9.5), 50 μm (E10.5), 60 μm (E11.5), 120 μm (E13.5) and 200 μm (E16.5).



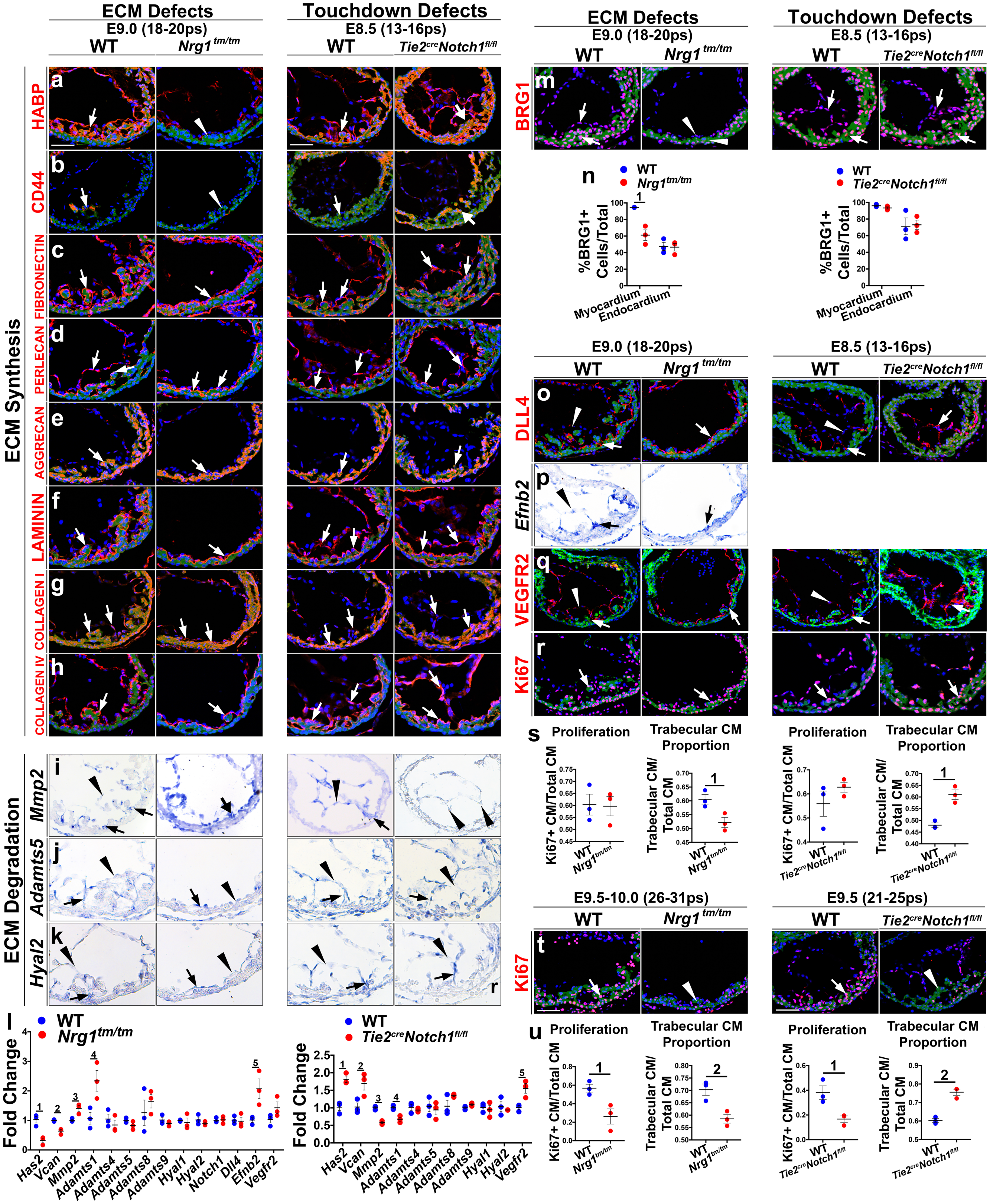
Extended Data Fig. 4 | See next page for caption.

Extended Data Fig. 4 | Morphological and molecular analysis of mutants with trabeculation defects. Related to Figs. 1–3. Histological and marker analysis of left ventricle in mutant embryos and somite-matched wild-type control embryos at indicated stages. **a**–**d**, Analysis of wild-type and *Erbb2*$^{-/-}$ mutants showing histology (**a**), normalized trabecular and ECM area quantifications (**b**), immunofluorescence for N1ICD (**c**) and fold change in number of N1ICD^{+} cells (**d**). **e**, **f**, Immunofluorescence analysis of pERBB2 in *Erbb2*$^{-/-}$ (**e** (right)) and *Nrg1*$^{tm/tm}$ (**f** (right)) mutant embryos compared to wild-type (**e**, **f**). **g**–**j**, Analysis of wild-type and *Tie1*^{cre}*VEGFR2*$^{fl/fl}$ mutants showing histology (**g**), normalized tissue area quantifications (**h**) and immunofluorescence of VEGFR2 (**i**) and pAKT (**j**). **k**–**p**, Analysis of wild-type and mutant histology and normalized tissue area quantifications for the following strains: *Tie2*^{cre}*Rbpj*$^{fl/fl}$ (**k**, **l**), *Efnb2*$^{gfp/gfp}$ (**m**, **n**) and *Bmp10*$^{-/-}$ (**o**, **p**). **q**–**u**, Time-course analysis of histology using Alcian blue staining of *Nrg1*$^{tm/tm}$ (**q**–**u** (middle)) and *Tie2*^{cre}*Notch1*$^{fl/fl}$ (**q**–**u** (right)) embryos relative to somite-matched wild-type (**q**–**u** (left)) at indicated stages. In all H&E panels, trabecular myocardium (black arrows), ECM (white arrows). In all panels showing molecular expression: marker expression (arrows), reduced marker expression (arrowheads). In all immunofluorescence images: markers associated with endocardium or ECM (red); markers associated with myocardium (orange); myocardium is stained for SMA (green dulled mask); nuclei (DAPI, blue).The cardiac lumen is indicated by 'l'. Trabecular myocardium (black arrows), ECM (white arrows). **a**–**u**, $n = 3$ independent embryos per genotype. Quantitative data are shown as mean ± s.e.m. Two-sided Student's *t*-tests (without corrections for multiple comparisons) were used. Significant comparisons are indicated by numbers as follows. **b**, 1, $P = 0.018$; 2, $P = 0.047$. **d**, 1, $P = 0.034$. **h**, 1, $P = 0.04$; 2, $P = 0.033$. **l**, 1, $P = 0.035$. Scale bars, 20 µm (**q**), 25 µm (**r**), 30 µm (**a**, **c**, **e**–**g**, **i**–**k**, **m**, **o**, **s**) and 35 µm (**t**, **u**).



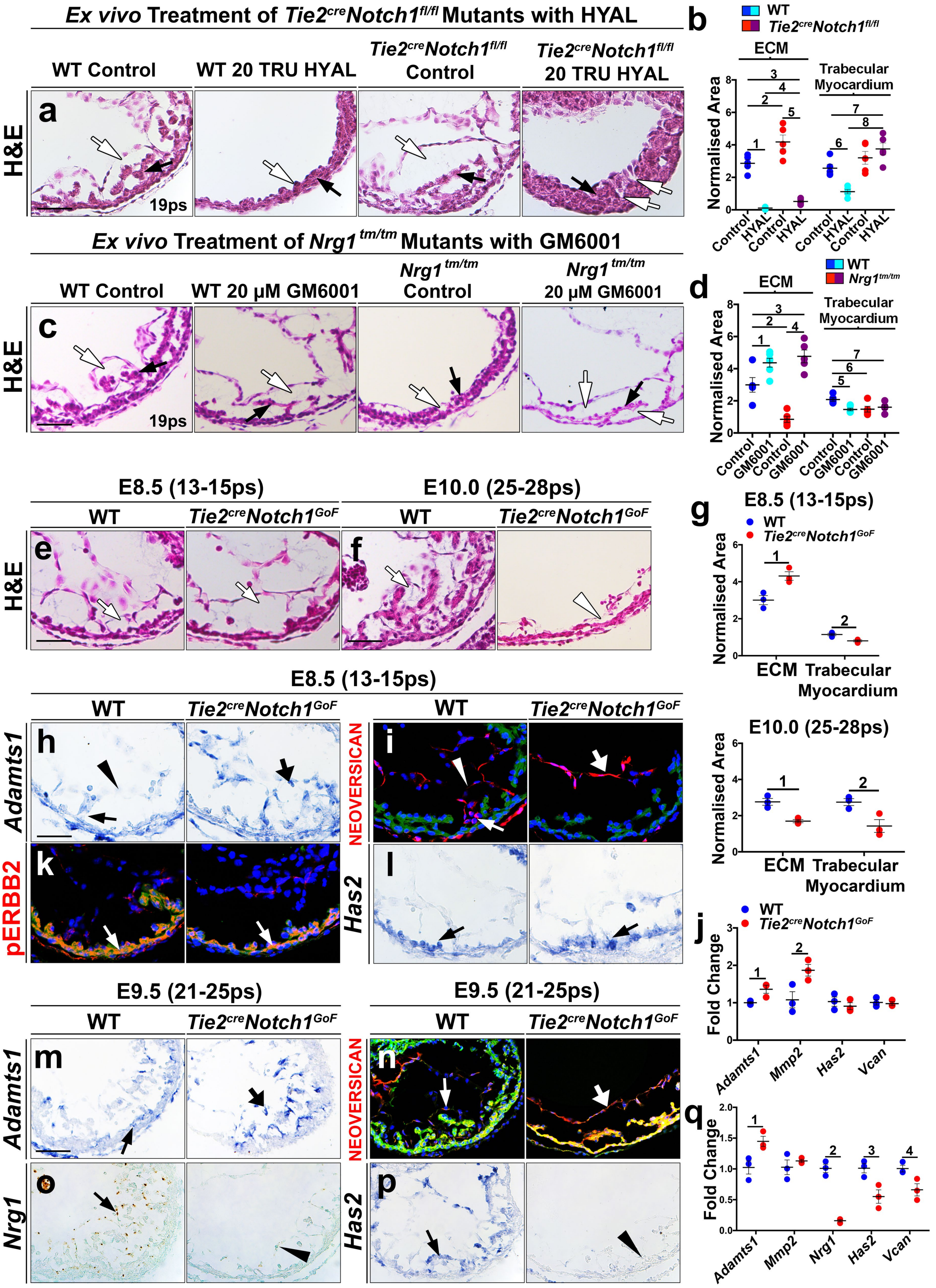
Extended Data Fig. 5 | See next page for caption.

Extended Data Fig. 5 | Marker analysis in *Nrg1*$^{tm/tm}$ and *Tie2*^{cre}*Notch1*$^{fl/fl}$ mutant strains. Related to Fig. 2. **a**–**i**, Analysis of markers in left ventricle sections by ISH, immunofluorescence and qPCR in *Nrg1*$^{tm/tm}$ (**a**–**i** (second left)), *Tie2-Notch1*$^{fl/fl}$ (**a**–**i** (right)) mutant embryos and somite-matched wild-type embryo (**a**–**h** (left, second from the right)) at indicated optimal time points. Data are organized around markers for chamber myocardium (*Mest*, *Nppa*), trabecular myocardium (*Bmp10*, *Gja5*, *Hopx*) and compact myocardium (*Hey2*, *Mycn*) as indicated. ISH (**a**–**h**); qPCR (**i**). **i**, Panels show qPCR analysis of the genes described in **a**–**h** in *Nrg1*$^{tm/tm}$ mutants at E9.0 and E10.0 (**i** (left column)) and in *Tie2*^{cre}*Notch1*$^{fl/fl}$ mutants at E8.5 and E9.5 (**i** (right column)), the latter time points in each case being beyond the phenotypic onset stages for each mutant. In all panels showing gene/protein expression: marker expression (arrow), reduced marker expression (arrowhead), increased marker expression (thick arrow). **a**–**h**, $n = 3$ independent embryos per genotype. **i**, $n = 3$ biological replicates each evaluated for three experimental replicates. Quantitative data are shown as mean ± s.e.m. Two-sided Student's *t*-tests (without corrections for multiple comparisons) were used. Significant comparisons are indicated by numbers as follows. **i**, Top left, 1, $P = 0.013$; bottom left, 1, $P = 0.0015$; 2, $P = 0.018$; 3, $P = 0.002$; 4, $P = 0.0027$; 5, $P = 0.048$; 6, $P = 0.0044$; bottom right, 1, $P = 0.04$; 2, $P = 0.0098$; 3, $P = 0.047$; 4, $P = 0.012$; 5, $P = 0.022$; 6, $P = 0.0029$; 7, $P = 0.02$. Scale bars, 30 μm (**a**–**h**).



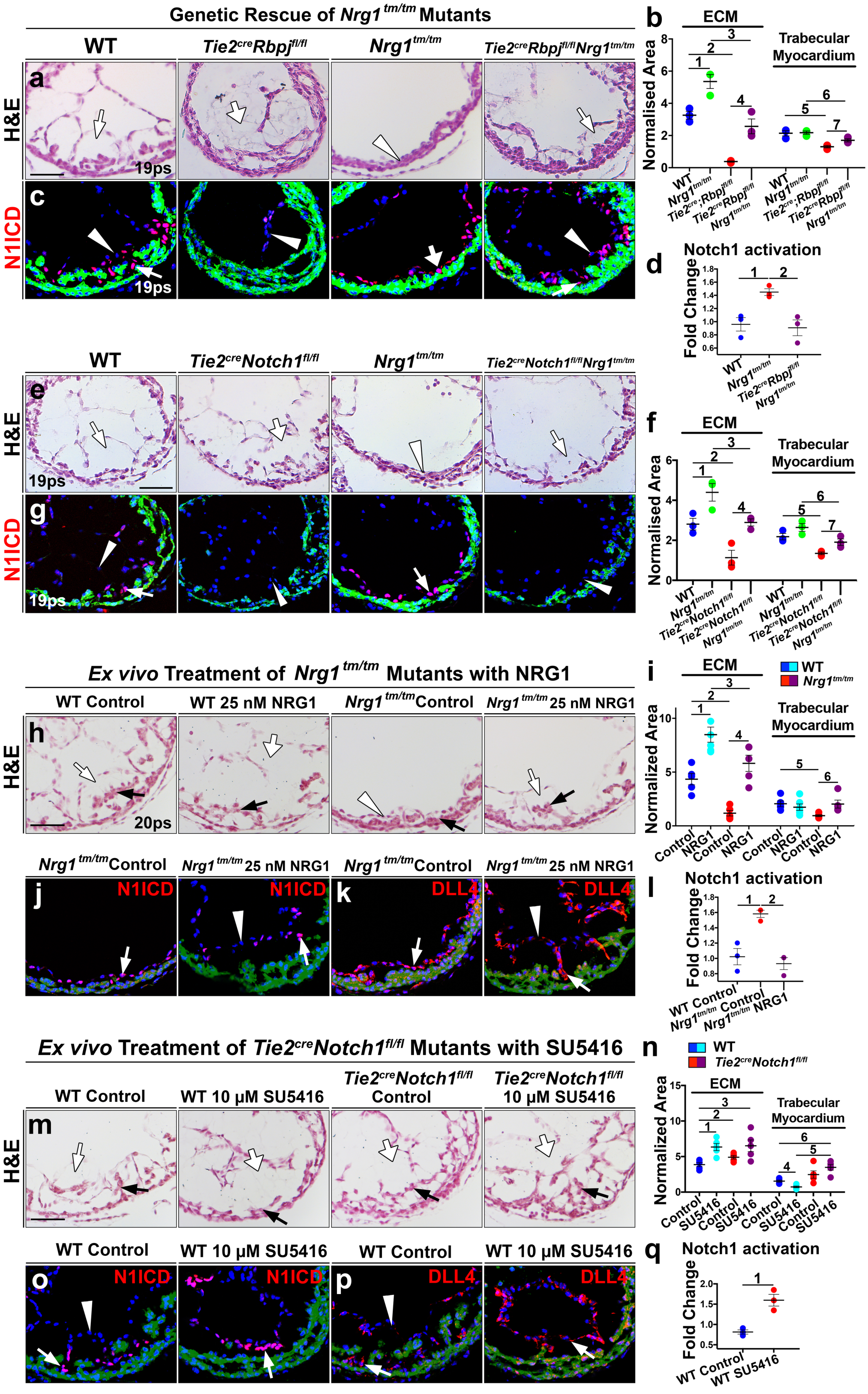
Extended Data Fig. 6 | See next page for caption.

Extended Data Fig. 6 | Analysis of ECM synthesis and degradation markers, and proliferation, in *Nrg1*$^{tm/tm}$ and *Tie2*^{cre}*Notch1*$^{fl/fl}$ mutants. Related to Fig. 2. **a**–**k**, **m**, **o**–**q**, Analysis of ECM synthesis (**a**–**h**) and degradation (**i**–**k**) markers, BRG1 (**m**), Notch pathway (**o**–**p**) and VEGFA pathway (**q**) in left ventricle of *Nrg1*$^{tm/tm}$ mutants at E9.0 (**a**–**q** (second left)) and *Tie2*^{cre}*Notch1*$^{fl/fl}$ mutants at E8.5 (**a**–**q** (right)) relative to wild-type embryos (**a**–**q** (left, second from the right)) showing immunofluorescence analysis of HABP (**a**), CD44 (**b**), fibronectin (**c**), perlecan (**d**), aggrecan (**e**), laminin (**f**), collagen type I (**g**), collagen type IV (**h**), BRG1 (**m**), DLL4 (**o**) and VEGFR2 (**q**); ISH for *Mmp2* (**i**), *Adamts5* (**j**), *Hyal2* (**k**) and *Efnb2* (**p**); and qPCR (**l**). **n**, Percentage of BRG1 positive cells compared to the total number of cells in the left ventricle endocardium and myocardium of *Nrg1*$^{tm/tm}$(**n** (left)) and *Tie2*^{cre}*Notch1*$^{fl/fl}$ (**n** (right)) mutants. **r**–**u**, Proliferation analysis by Ki-67 immunofluorescence of *Nrg1*$^{tm/tm}$ mutants at E9.0 (**r** (second left)) and E9.5–E10.0 (**t** (second left)); *Tie2*^{cre}*Notch1*$^{fl/fl}$ mutants at E8.5 (**r** (right)) and E9.5 (**t** (right)), relative to wild-type embryos (**r**, **t**, (left, second from the right); and the related quantifications of total myocardium proliferation and the proportion of trabecular cardiomyocytes compared to the total cardiomyocytes (**s**, **u**). Gene/protein expression panels: expression (arrow), reduced expression (arrowhead). In all immunofluorescence: markers associated with endocardium or ECM (red); markers associated with myocardium (orange); myocardium is stained for SMA (green dulled mask); nuclei (DAPI, blue). **a**–**u**, $n = 3$ independent embryos per genotype. **l**, $n = 3$ biological replicates each evaluated for three experimental replicates. Quantitative data are shown as mean ± s.e.m. Two-sided Student's *t*-tests (without corrections for multiple comparisons) were used. Significant comparisons are indicated by numbers as follows. **l**, Left, 1, $P = 0.0066$; 2, $P = 0.0038$; 3, $P = 0.017$, 4, $P = 0.035$; 5, $P = 0.036$; right, 1, $P = 0.0039$; 2, $P = 0.039$; 3, $P = 0.00053$; 4, $P = 0.0074$; 5, $P = 0.022$. **n**, 1, $P = 0.006$. **s**, Left, 1, $P = 0.029$; right, 1, $P = 0.012$. **u**, Left, 1, $P = 0.045$; 2, $P = 0.017$; right, 1, $P = 0.04$; 2, $P = 0.0031$. Scale bars, 55 μm (**a**–**k**, **m**, **o**–**r**) and 40 μm (**t**).



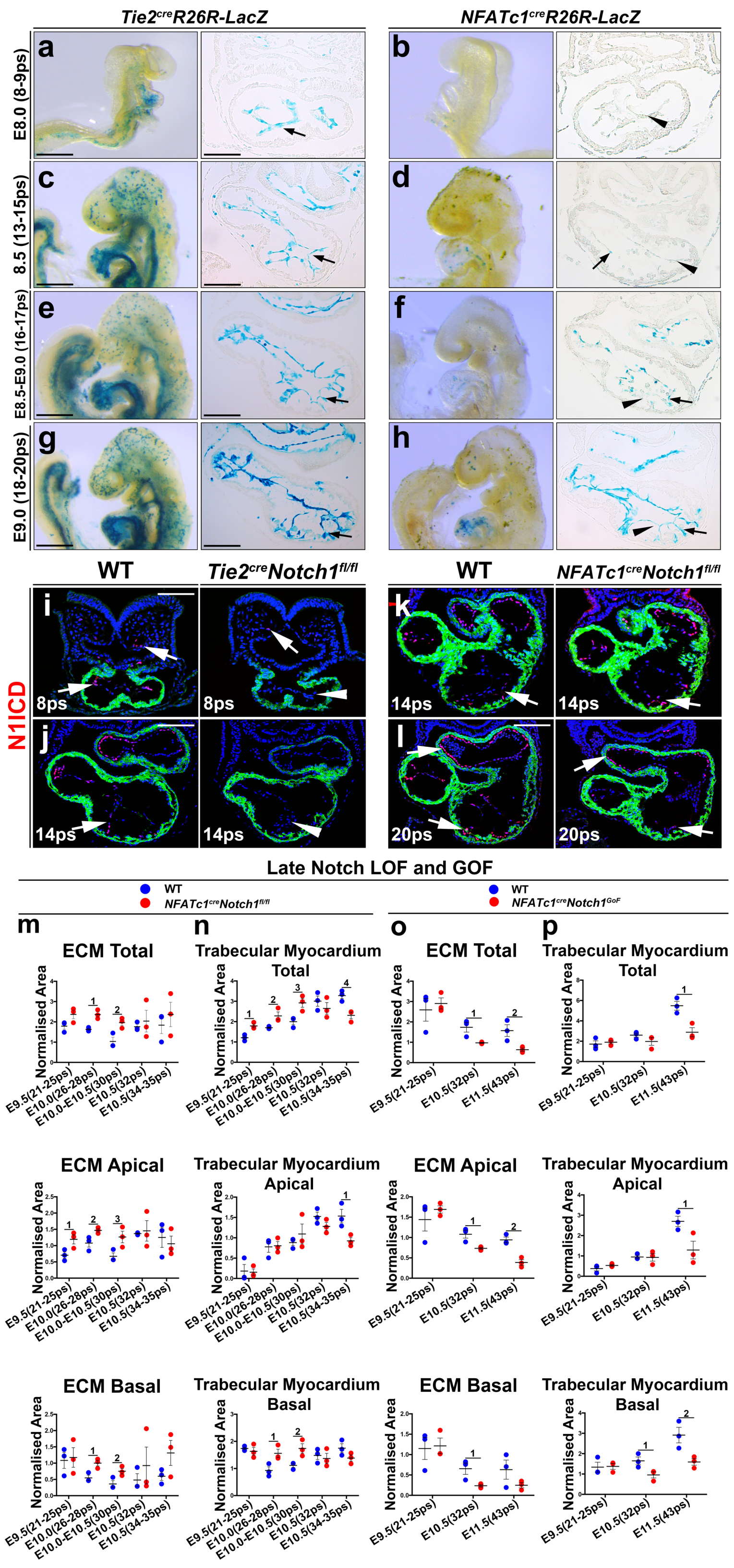
Extended Data Fig. 7 | See next page for caption.

Extended Data Fig. 7 | Manipulation of ECM dynamics in RBEC assays and marker analysis of $Tie2^{cre}Notch1^{GOF}$ mutant embryos. Related to Fig. 2. **a**–**d**, H&E analysis of wild-type and $Tie2^{cre}Notch1^{fl/fl}$ embryos treated with 20 turbidity reducing units (TRU) of hyaluronidase (HYAL) or PBS as control (**a**, **b**) or wild-type and $Nrg1^{tm/tm}$ embryos treated with 20 µM GM6001 or DMSO (**c**, **d**) in RBECs for 24 h, and accompanying ECM and trabecular area quantifications normalized to chamber perimeter (**b**, **d**). **e**–**g**, H&E of $Tie2^{cre}Notch^{GOF}$ embryos (**e**, **f** (right)) compared to wild-type embryos (**e**, **f** (left)) at E8.5 and E10.0 with morphological quantification of ECM and trabecular areas normalized to chamber perimeter (**g**). **h**–**q**, Molecular analysis of $Tie2^{cre}Notch^{GOF}$ embryos relative to wild-type embryos at E8.5 (**h**–**l**) and E9.5 (**m**–**q**): ISH of *Adamts1* (**h**, **m**) and *Has2* (**l**, **p**); immunofluorescence analysis of neoversican (**i**, **n**) and pERBB2 (**k**); RNA-scope of *Nrg1* (**o**); qPCR analysis of the respective markers (**j**, **q**). Gene/protein expression panels: expression (arrow), reduced expression (arrowhead), increased expression (thick arrow). In all immunofluorescence images: markers associated with endocardium or ECM (red); markers associated with myocardium (orange); myocardium is stained for SMA (green dulled mask); nuclei (DAPI, blue). **e**–**i**, **k**–**p**, $n = 3$ independent embryos were used; **a**–**d**, $n = 5$, except in **d**, $n = 6$ wild-type embryos treated with GM6001. **j**, **q**, $n = 3$ biological replicates each evaluated for three experimental replicates. Quantitative data are shown as mean ± s.e.m. Two-sided Student's *t*-tests (without corrections for multiple comparisons) were used. Significant comparisons are indicated by numbers as follows. **b**, 1, $P = 1.99 \times 10^{-6}$; 2, $P = 0.025$; 3, $P = 9.28 \times 10^{-6}$; 4, $P = 0.00085$; $5 = 2.83 \times 10^{-5}$; 6, $P = 0.00067$; 7, $P = 0.026$; 8, $P = 0.00014$. **d**, 1, $P = 0.024$; 2, $P = 0.0024$; 3, $P = 0.018$; 4, $P = 0.000021$; 5, $P = 0.016$; 6, $P = 0.022$; 7, $P = 0.028$. **g**, Top, 1, $P = 0.018$; 2, $P = 0.017$; bottom, 1, $P = 0.0071$; 2, $P = 0.03$. **j**, 1, $P = 0.03$; 2, $P = 0.041$. **q**, 1, $P = 0.034$; 2, $P = 0.00026$; 3, $P = 0.025$; $4 = 0.038$. Scale bars, 55 µm (**a**, **c**), 30 µm (**e**, **h**–**l**) and 40 µm (**m**–**p**).



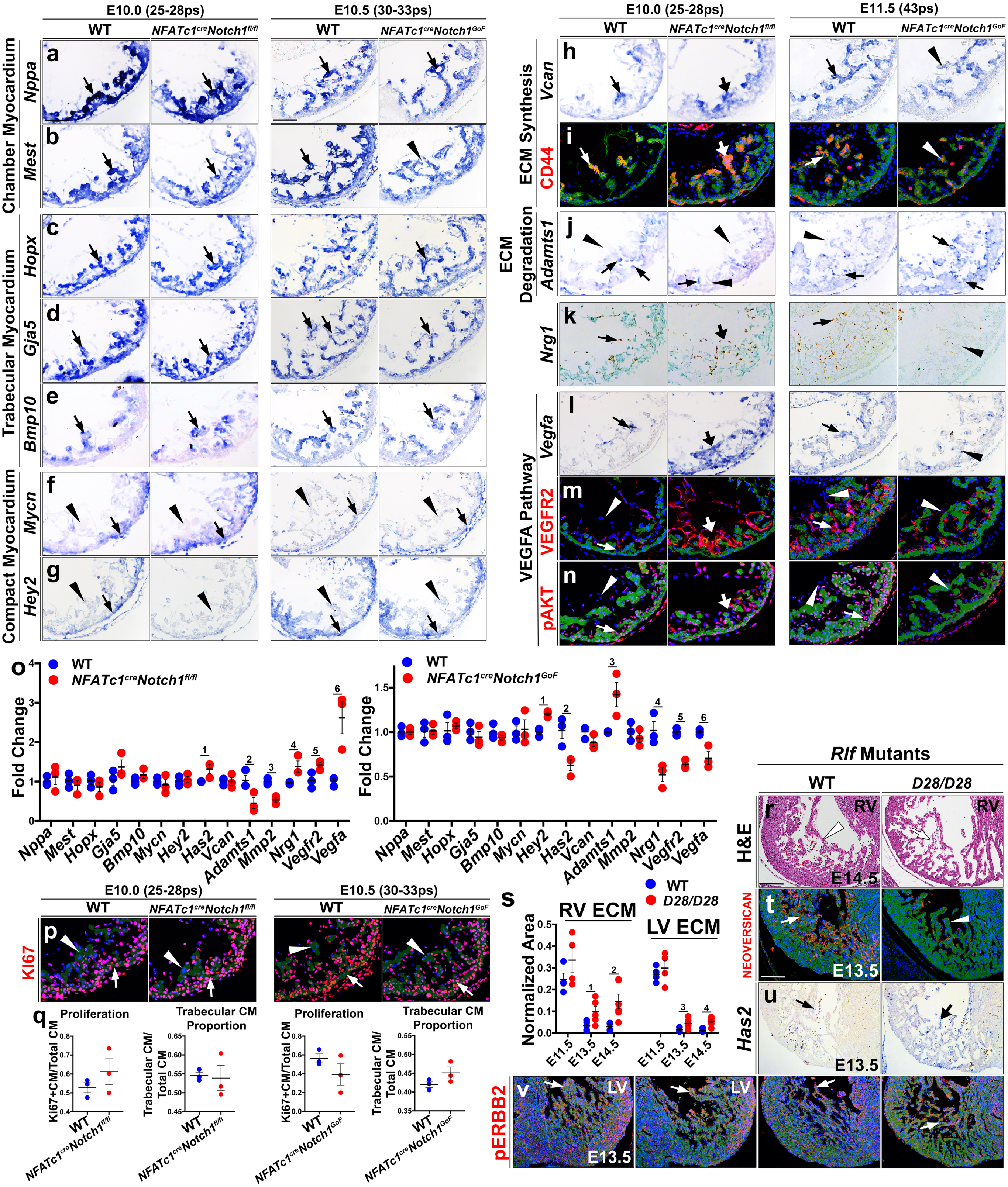
Extended Data Fig. 8 | See next page for caption.

Extended Data Fig. 8 | Genetic and chemical mutant rescue assays. Related to Figs. 2, 3. **a**–**d**, Analysis of wild-type, *Nrg1*$^{tm/tm}$, *Tie2*^{cre}*Rbpj*$^{fl/fl}$ and *Tie2*^{cre}*Rbpj*$^{fl/fl}$*Nrg1*$^{tm/tm}$ mutant embryos showing histology (**a**), normalized ECM and trabecular area quantifications (**b**), immunofluorescence of N1ICD (**c**) and fold change in the number of N1ICD^{+} cells (**d**). **e**–**g**, Analysis of wild-type, *Nrg1*$^{tm/tm}$, *Tie2*^{cre}*Notch1*$^{fl/fl}$ and *Tie2*^{cre}*Notch1*$^{fl/fl}$*Nrg1*$^{tm/tm}$ mutant embryos showing histology (**e**), normalized ECM and trabecular area quantifications (**f**) and immunofluorescence of N1ICD (**g**). Note the persistent N1ICD staining in the *Tie2*^{cre}*Rbpj*$^{fl/fl}$*Nrg1*$^{tm/tm}$ mutant due to mosaic Cre-mediated deletion of *Notch1*. Here, N1ICD appears to nucleate touchdowns. Cre-mediated deletion was complete in *Tie2*^{cre}*Notch1*$^{fl/fl}$*Nrg1*$^{tm/tm}$ mutants at E9.0, although mosaicism in Cre-recombination at an earlier time point may allow N1ICD restriction and nucleation of touchdowns. **h**, **i**, H&E of *Nrg1*$^{tm/tm}$ and wild-type embryos treated with 25 nM NRG1 or PBS as control for 24 h (**h**) with associated area quantifications (**i**). **j**–**l**, Immunofluorescence of N1ICD (**j**) and DLL4 (**k**) with quantification the number of N1ICD^{+} cells (**l**). **m**–**q**, Analysis of wild-type and *Tie2*^{cre}*Notch1*$^{fl/fl}$ embryos treated with 10 μM SU5416 or DMSO control for 24 h, showing histology (**m**), ECM and trabecular area quantifications (**n**) and immunofluorescence in SU5416- and control-treated wild-type embryos of N1ICD (**o**) and DLL4 (**p**), with fold change in the number of N1ICD^{+} cells (**q**). In H&E panels: trabecular myocardium (black arrows), ECM (white arrows). In all panels showing molecular expression: expression (arrows), reduced expression (arrowheads), increased expression (thick arrows). In all immunofluorescence images: markers associated with endocardium or ECM (red); markers associated with myocardium (orange); myocardium is stained for SMA (green dulled mask); nuclei (DAPI, blue). **a**–**f**, **j**–**l**, **o**–**q**, $n = 3$ independent embryos were used; **h**, **i**, **m**, **n**, $n = 5$, except in **h**, **i**, $n = 6$ wild-type embryos treated with NRG1. Quantitative data are shown as mean ± s.e.m. Two-sided Student's *t*-tests (without corrections for multiple comparisons) were used. Significant comparisons are indicated by numbers as follows. **b**, 1, $P = 0.013$; 2, $P = 0.0003$; 3, $P = 0.012$; 4, $P = 0.009$; 5, $P = 0.009$; 6, $P = 0.029$; 7, $P = 0.033$. **d**, 1, $P = 0.024$; 2, $P = 0.03$. **f**, 1, $P = 0.039$; 2, $P = 0.023$; 3, $P = 0.034$; 4, $P = 0.012$; 5, $P = 0.009$; 6, $P = 0.045$; 7, $P = 0.034$. **i**, 1, $P = 0.0014$; 2, $P = 0.0006$; 3, $P = 0.03$; 4, $P = 0.0004$; 5, $P = 0.0046$; 6, $P = 0.022$. **l**, 1, $P = 0.02$; 2, $P = 0.0042$. **n**, 1, $P = 0.0025$; 2, $P = 0.024$; 3, $P = 0.016$; 4, $P = 0.0033$; 5, $P = 0.00015$; 6, $P = 0.0017$. **q**, 1, $P = 0.022$. Scale bars, 50 μm (**a**, **c**, **e**, **g**, **h**, **j**, **k**, **m**, **o**, **p**).



Extended Data Fig. 9 | See next page for caption.

Extended Data Fig. 9 | Analysis of deletion efficiency of the *R26R-LacZ* Cre-reporter using *Tie2*^{cre} and *Nfatc1*^{cre} driver lines and quantification analysis of late NOTCH1 loss-of-function and gain-of-function mutants. Related to Figs. 1–3. **a**–**h**, Time-course analysis of staining of β-galactosidase in whole-mount embryos and representative sections from *Tie2*^{cre}*R26R-LacZ* and *Nfatc1*^{cre}*R26R-LacZ* embryos at indicated stages. **i**–**l**, Immunofluorescence analysis of N1ICD in *Tie2*^{cre}*Notch1*$^{fl/fl}$ embryos at 8 ps and 14 ps, and in *Nfatc1*^{cre}*Notch1*$^{fl/fl}$ embryos at 14 ps and 20 ps, compared to wild-type embryos, as indicated. **m**–**p**, Quantification of normalized ECM and trabecular myocardium areas measuring the total, apical or basal trabecular regions in *Nfatc1*^{cre}*Notch1*$^{fl/fl}$ (**m**, **n**) and *Nfatc1*^{cre}*Notch1*^{GOF} (**o**, **p**) mutant embryos. In all panels showing gene/protein expression: marker (arrow), reduced expression (arrowhead), increased expression (thick arrow). In all immunofluorescence images: markers associated with endocardium or ECM (red); markers associated with myocardium (orange); myocardium stained for SMA (green dulled mask); nuclei (DAPI, blue). **a**–**p**, $n = 3$ independent embryos were used. Quantitative data are shown as mean ± s.e.m. Two-sided Student's *t*-tests (without corrections for multiple comparisons) were used. Significant comparisons are indicated by numbers as follows. **m**, Top, 1, $P = 0.0057$; 2, $P = 0.021$; middle, 1, $P = 0.048$; 2, $P = 0.045$; 3, $P = 0.046$; bottom, 1, $P = 0.018$; 2, $P = 0.041$. **n**, Top, 1, $P = 0.0099$; 2, $P = 0.043$; 3, $P = 0.025$; 4, $P = 0.012$; middle, 1, $P = 0.03$; bottom, 1, $P = 0.036$; 2, $P = 0.030$. **o**, Top, 1, $P = 0.032$; 2, $P = 0.035$; middle, 1, $P = 0.025$; 2, $P = 0.0048$; bottom, 1, $P = 0.042$. **p**, Top, 1, $P = 0.014$; middle, 1, $P = 0.049$; bottom, 1, $P = 0.048$; 2, $P = 0.034$. Scale bars, 250 μm (**a**, **b** (left)), 50 μm (**a**, **b** (right)), 340 μm (**c**, **d** (left)), 65 μm (**c**, **d** (right)), 360 μm (**e**, **f** (left)), 70 μm (**e**, **f** (right)), 430 μm (**g**, **h** (left)), 80 μm (**g**, **h** (right), **i**–**k**) and 100 μm (**l**).



Extended Data Fig. 10 | See next page for caption.

Extended Data Fig. 10 | Marker analysis in *Nfatc1^{cre}Notch1$^{fl/fl}$*, *Nfatc1^{cre}Notch1^{GOF}* and *Rlf$^{D28/D28}$* mutant strains. Related to Fig. 3. **a**–**p**, Analysis of markers in left ventricle sections of *Nfatc1^{cre}Notch1$^{fl/fl}$* (**a**–**p** (second left)) and *Nfatc1^{cre}Notch1^{GOF}* (**a**–**p** (right)) mutant embryos and somite-matched wild-type embryos (**a**–**p** (left, second from the right)). Data are organized around markers for chamber myocardium (*Mest*, *Nppa*), trabecular myocardium (*Bmp10*, *Gja5*, *Hopx*), compact myocardium (*Hey2*, *Mycn*), ECM synthesis (CD44, *Vcan*), ECM degradation (*Adamts1*), NRG1 pathway (*Nrg1*) and VEGFA pathway (*Vegfa*, VEGFR2, pAKT), and proliferation (Ki-67) as indicated. ISH (**a**–**h**, **j**, **l**); RNA-scope (**k**); immunofluorescence (**i**, **m**, **n**, **p**). **o**, qPCR analysis of the genes described in **a**–**n** in *Nfatc1^{cre}Notch1$^{fl/fl}$* mutants at E10.0 (left) and *Nfatc1^{cre}Notch1^{GOF}* mutants at E10.5 (right). **q**, Quantification of myocardial proliferation as the ratio of the number of Ki-67 positive cardiomyocytes versus the total number of cardiomyocytes and the proportion of trabecular cardiomyocytes out of the total number of cardiomyocytes in left ventricle from *Nfatc1^{cre}Notch1$^{fl/fl}$* mutants at E10.0 (left two panels) and *Nfatc1^{cre}Notch1^{GOF}* mutants at E10.5 (right two panels). **r**–**v**, Analysis of wild-type and *Rlf$^{D28/D28}$* mutants showing right ventricle (RV) histology at E14.5 (**r**), normalized ECM area quantification in both ventricles (**s**), immunofluorescence of neoversican in the right ventricle (**t**) and pERBB2 in both ventricles (**v**), and ISH of *Has2* in the right ventricle (**u**) at E13.5. In all panels showing gene/protein expression: expression (arrow), reduced expression (arrowhead), increased expression (thick arrow). In all immunofluorescence images: markers associated with endocardium or ECM (red); markers associated with myocardium (orange); myocardium is stained for SMA (green dulled mask); nuclei (DAPI, blue). **a**–**n**, **p**–**q**, **t**–**v**, $n = 3$ independent embryos were used; s, $n = 4$ E11.5 and wild-type E14.5, $n = 5$ E13.5, $n = 6$ *Rlf$^{D28/D28}$* E14.5. **o**, $n = 3$ biological replicates each evaluated for three experimental replicates. Quantitative data are shown as mean ± s.e.m. Two-sided Student's *t*-tests (without corrections for multiple comparisons) were used. Significant comparisons are indicated by numbers as follows. **o**, Left, 1, $P = 0.047$; 2, $P = 0.04$; 3, $P = 0.0029$; 4, $P = 0.044$; 5, $P = 0.041$; 6, $P = 0.017$; right, 1, $P = 0.0037$; 2, $P = 0.025$; 3, $P = 0.036$; 4, $P = 0.017$; 5, $P = 0.00057$; 6, $P = 0.017$. **s**, 1, $P = 0.045$; 2, $P = 0.026$; 3, $P = 0.042$; $4 = 0.0033$. Scale bars, 50 μm (**a**–**n**, **p**, unless otherwise indicated), 60 μm (**h**–**n**, right two panels) 85 μm (**t**–**v**) and 150 μm (**r**).

LETTER

https://doi.org/10.1038/s41586-018-0022-5

Vms1 and ANKZF1 peptidyl-tRNA hydrolases release nascent chains from stalled ribosomes

Rati Verma[1,2,3], Kurt M. Reichermeier[2,5], A. Maxwell Burroughs[4], Robert S. Oania[1,2], Justin M. Reitsma[2], L. Aravind[4]* & Raymond J. Deshaies[1,2,3]*

Ribosomal surveillance pathways scan for ribosomes that are transiently paused or terminally stalled owing to structural elements in mRNAs or nascent chain sequences[1, 2]. Some stalls in budding yeast are sensed by the GTPase Hbs1, which loads Dom34, a catalytically inactive member of the archaeo-eukaryotic release factor 1 superfamily. Hbs1–Dom34 and the ATPase Rli1 dissociate stalled ribosomes into 40S and 60S subunits. However, the 60S subunits retain the peptidyl-tRNA nascent chains, which recruit the ribosome quality control complex that consists of Rqc1–Rqc2–Ltn1–Cdc48–Ufd1–Npl4. Nascent chains ubiquitylated by the E3 ubiquitin ligase Ltn1 are extracted from the 60S subunit by the ATPase Cdc48–Ufd1–Npl4 and presented to the 26S proteasome for degradation[3–9]. Failure to degrade the nascent chains leads to protein aggregation and proteotoxic stress in yeast and neurodegeneration in mice[10–14]. Despite intensive investigations on the ribosome quality control pathway, it is not known how the tRNA is hydrolysed from the ubiquitylated nascent chain before its degradation. Here we show that the Cdc48 adaptor Vms1 is a peptidyl-tRNA hydrolase. Similar to classical eukaryotic release factor 1, Vms1 activity is dependent on a conserved catalytic glutamine. Evolutionary analysis indicates that yeast Vms1 is the founding member of a clade of eukaryotic release factor 1 homologues that we designate the Vms1-like release factor 1 clade.

Yeast Cdc48, an AAA + (ATPases associated with diverse cellular functions) ATPase that is conserved across eukaryotes and archaea, is a protein unfoldase[15, 16] that engages in myriad cellular functions through the binding of adaptors such as Ufd1–Npl4, which bind both Cdc48 and ubiquitylated substrate proteins. A protein A-based 'non-stop' substrate reporter (PrANS) encoded by an mRNA lacking a stop codon (Fig. 1b) accumulates on the ribosomal 60S subunit as a ubiquitylated species linked to tRNA when Cdc48 or Ufd1–Npl4 function is compromised[7], suggesting that Cdc48–Ufd1–Npl4 somehow extracts the ubiquitylated nascent chain from the 60S tunnel[6–8].

Two pieces of evidence implicated Vms1, which contains a Cdc48-binding VIM motif[17, 18], in the ribosome quality control (RQC) pathway. First, *vms1*Δ and RQC-deficient mutations cause a synthetic growth defect when combined with mutations that impair degradation of non-stop mRNA[8, 19]. Second, *vms1*Δ[18] mutants, like *rqc2*Δ[20], are sensitive to cycloheximide treatment. We extended this by showing that *vms1*Δ, like *ltn1*Δ *(LTN1* is also known as *RKR1)* and *cdc48*[7], displayed sensitivity to hygromycin, which binds the decoding centre (Extended Data Fig. 1a).

To investigate further the role of Vms1 in the degradation of non-stop proteins, we examined PrANS levels. PrANS conjugated to tRNA (PrANS–tRNA) accumulated in *vms1*Δ cells, but not in mutants lacking other Cdc48 adaptors (Fig. 1c). Additionally, *vms1*Δ cells accumulated high molecular mass forms of PrANS. Treatment with RNase or deubiquitylase enzymes (Extended Data Fig. 1c) and pull-downs with ubiquitin-binding TUBE resin (Extended Data Fig. 1d) confirmed that high molecular mass PrANS species were tRNA-linked ubiquitin conjugates (Ub–PrANS–tRNA).

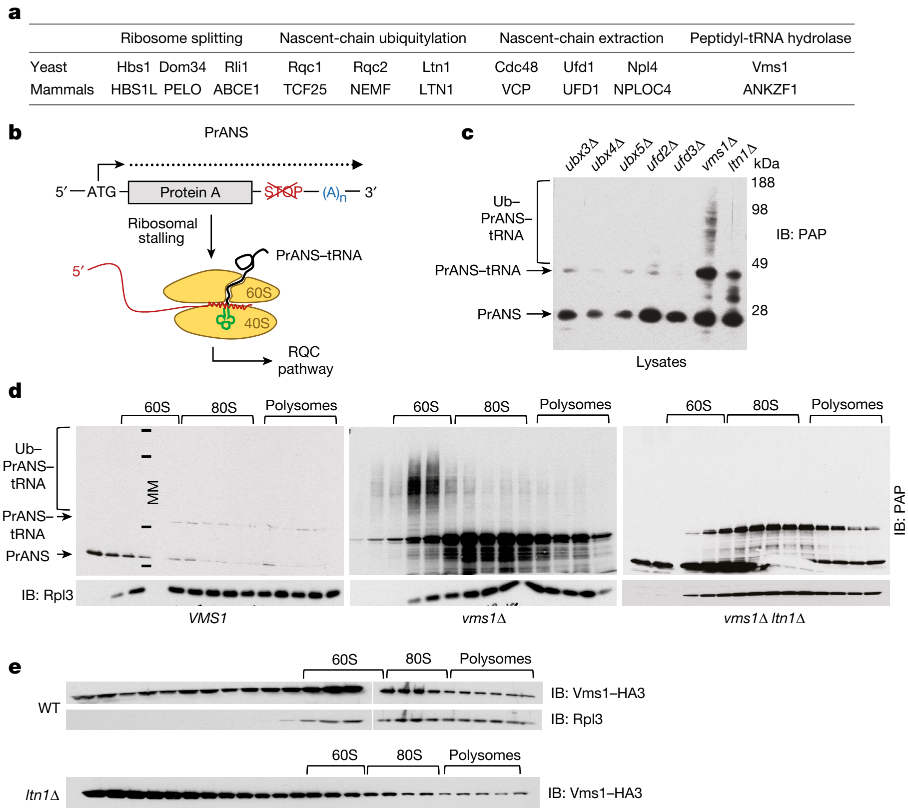
Splitting of stalled 80S ribosomes is a prerequisite for the stable association of Ltn1 with 60S subunits. Ltn1 then ubiquitylates the nascent chains before their degradation by 26S proteasomes[21]. To determine whether Ub–PrANS–tRNA accumulated on 60S subunits in *vms1*Δ cells, we fractionated cell lysates on sucrose gradients. Immunoblotting for PrANS and the 60S protein Rpl3 revealed that Ub–PrANS–tRNA accumulated maximally in the first few Rpl3-containing fractions, which marks the position of 60S subunits[7] (Fig. 1d). By contrast, non-ubiquitylated PrANS–tRNA that accumulated in both *vms1*Δ and *vms1*Δ*ltn1*Δ cells was enriched on 80S subunits and polysomes. These data confirm that Ub–PrANS–tRNA accumulated on 60S subunits in *vms1*Δ in a manner that was dependent on Ltn1.

If Vms1 is directly involved in releasing stalled PrANS, it should bind ribosomes. To investigate this prediction, we immunoblotted sucrose gradient fractions from cells expressing Vms1 tagged with a triple haemagglutinin epitope (Vms1–HA3). Ltn1 is restricted to 60S fractions[4], whereas Vms1–HA3 was detected across the entire gradient (Fig. 1e). This behaviour suggests a broader function for Vms1 beyond its role in the RQC pathway. Notably, the 60S peak of Vms1–HA3 was diminished and shifted towards lighter fractions in *ltn1*Δ cells (Fig. 1e), suggesting that in addition to Cdc48–Ufd1–Npl4[6], Ltn1 also promotes the recruitment of Vms1 to 60S subunits. However, unlike Cdc48–Ufd1–Npl4, Vms1 was not identified as a component of the RQC complex[6]. Consistent with this, Vms1 did not associate directly with Ufd1–Npl4 (Extended Data Fig. 1e–h).

Given the exceptionally strong accumulation of PrANS–tRNA in *vms1*Δ mutants, we wondered whether Vms1 might have any relationship to known peptidyl-tRNA hydrolases. We carried out sensitive iterative sequence profile searches with PSI-BLAST and hidden Markov models seeded with the core RNase H fold domain of the archaeal and eukaryotic release factor 1 (RF1) proteins and the catalytically inactive Dom34 proteins, which are both members of the archaeo-eukaryotic RF1 (aeRF1) superfamily (see 'Computational analyses' in Methods). Notably, these searches recovered the central globular region of Vms1 orthologues. Reciprocal PSI-BLAST searches with this region recovered several members of the aeRF1 superfamily. This relationship was confirmed with a profile–profile comparison using the HHpred program with an alignment of the central conserved region of Vms1 run against the Protein Data Bank (PDB) database, which significantly recovered human eRF1 (PDB accession number 1DT9; $P = 4 \times 10^{-5}$). This showed that the central domain of Vms1 had a putative catalytic loop bearing a conserved glutamine comparable to the classic active site of aeRF1 proteins (Fig. 2a, b). An aeRF1-like topology was observed in the recently reported crystal structure of the middle domain of Vms1, but the putative active-site loop was not resolved[22]. The conserved glutamine in the eRF1 protein coordinates a water molecule in the A site at the peptidyl transferase centre (PTC). Nucleophilic attack of the peptidyl-tRNA in the P site by this water molecule releases the completed polypeptide chain[23, 24]. In addition to the glutamine in the putative active site, Vms1 also contains conserved vicinal arginines

---

[1]Howard Hughes Medical Institute, California Institute of Technology, Pasadena, CA, USA. [2]Division of Biology and Biological Engineering, California Institute of Technology, Pasadena, CA, USA. [3]Amgen Discovery Research, Thousand Oaks, CA, USA. [4]National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD, USA. [5]Present address: Genentech, South San Francisco, CA, USA. *e-mail: aravind@ncbi.nlm.nih.gov; rdeshaie@amgen.com



	Ribosome splitting			Nascent-chain ubiquitylation			Nascent-chain extraction			Peptidyl-tRNA hydrolase
Yeast	Hbs1	Dom34	Rli1	Rqc1	Rqc2	Ltn1	Cdc48	Ufd1	Npl4	Vms1
Mammals	HBS1L	PELO	ABCE1	TCF25	NEMF	LTN1	VCP	UFD1	NPLOC4	ANKZF1

Fig. 1 | Vms1 is required for release of ubiquitylated, tRNA-linked non-stop protein A reporter (Ub–PrANS–tRNA) from 60S ribosomal subunits. a, Primary components of the yeast RQC pathway and their mammalian orthologues. VCP is also known as p97. **b**, Model substrate reporter used to study the non-stop pathway. **c**, Mutants were lysed in denaturing SDS, fractionated on a NuPAGE gel and immunoblotted (IB) with peroxidase anti-peroxidase soluble complex (PAP) to detect protein A. *UFD3* is also known as *DOA1*. **d**, Native lysates (10 A_{260} units) were separated on sucrose gradients and fractions were resolved on Tris–glycine gels and immunoblotted with PAP and an anti-Rpl3 antibody. Molecular masses (MM) 250, 150, 50 and 23 kDa are indicated by marks overlaid on the gel. **e**, Native lysates from wild-type (WT) and *ltn1Δ* cells expressing Vms1–HA3 were fractionated on sucrose gradients and immunoblotted with anti-haemagglutinin and anti-Rpl3 antibodies. All western blots are representative of two biological replicates. Gel source images are provided in Supplementary Fig. 1.

in the α-helix immediately following the active-site loop (Fig. 2a, b). Equivalent arginines in eRF1 have been implicated in the stabilization of the active-site loop[23], and stimulation of GTPase eRF3[25]. Based on these similarities, we designated Vms1 as the founding member of the Vms1-like release factor 1 (VLRF1) clade within the aeRF1 superfamily. In addition to the eukaryotic orthologues of Vms1, the VLRF1 clade also contains members from certain bacteria, predominantly bacteroidetes and more divergent versions in certain Archaea (Extended Data Fig. 2a, b).

To verify that Vms1 is a novel peptidyl-tRNA hydrolase we generated two separate point mutations (Q295L and Q295P) of the predicted catalytic glutamine, which abrogate eRF1 function in budding yeast[23] and bacterial release factors[26], respectively. We also generated a putative active site deletion mutant (ΔAS). All mutant proteins were expressed normally and bound ribosomes (Fig. 3a). Whereas wild-type Vms1–HA3 rescued *vms1Δ*, cells expressing the point mutants accumulated PrANS-tRNA (Fig. 3a, left), which sedimented with ribosomes (Fig. 3a, middle) and was conjugated to ubiquitin (Fig. 3a, right). Sucrose gradient fractionation confirmed that Ub–PrANS–tRNA accumulated on 60S subunits in cells expressing Vms1-Q295L–HA3 (Fig. 3b). Notably, the mutant protein strongly accumulated on 60S subunits, and was depleted from low molecular mass fractions (Fig. 3c). Consistent with their biochemical defect, the $vms1^{Q295L}$, $vms1^{Q295P}$ and $vms1^{\Delta AS}$ mutants were extremely sensitive to hygromycin and cycloheximide (Fig. 3d). The point mutants thus phenocopy *vms1Δ*, both genetically and biochemically.

The putative catalytic region of Vms1 and other VLRF1 proteins is flanked by additional N- and C-terminal domains. The N-terminal C2H2-type zinc-finger of Vms1 is specifically related to those of Rei1 and certain SBDS paralogues, which function late in 60S subunit maturation[27]. This suggests that Vms1 might contact the ribosome with its Rei1-like zinc-finger domain (Fig. 2c). Meanwhile, the Vms1 C-terminal region is characterized by a run of three ankyrin repeats, a treble-clef fold zinc-binding domain (VTC; lost in fungi) and a VIM motif that binds Cdc48[17,18] (Fig. 2c). To probe the role of these other sequence features, we generated mutants including DNKR (lacking four Rei1-like zinc-finger residues predicted to interact with the 60S subunit; see legend to Extended Data Fig. 3c for details), RR (lacking the conserved vicinal arginines in the catalytic domain) and ΔVIM, and expressed the proteins in a *vms1Δ* strain. All mutant proteins were expressed comparably to wild type (Extended Data Fig. 3a). To assess the effects of the mutations, we performed sucrose gradient fractionation of ribosomes. Compared to *vms1Δ* rescued by wild-type $VMS1^{HA3}$, all mutants accumulated Ub–PrANS–tRNA in 60S fractions (Extended Data Fig. 3c). The RR mutant exhibited the strongest accumulation, as well as conspicuous spreading of Ub–PrANS–tRNA into the 80S peak, which was also observed with the ΔVIM mutant. Notably, the DNKR mutant protein was shifted to lighter fractions, with depletion from monosome and polysome fractions, consistent with a role for the zinc-finger in binding ribosomes (Extended Data Fig. 3d). Notably, the cycloheximide sensitivity of the various mutants mimics the data obtained with the reporter, with the ΔAS and RR mutants being most sensitive (Extended Data Fig. 3b).

To address conclusively the hypothesis that Vms1 is a novel peptidyl-tRNA hydrolase, we sought to reconstitute peptidyl-tRNA hydrolytic activity in vitro. Polyhistidine (His6) tags were added to the N-terminus of wild-type Vms1 and the Vms1-Q295L mutant. The proteins were then expressed in bacteria and purified (Extended Data

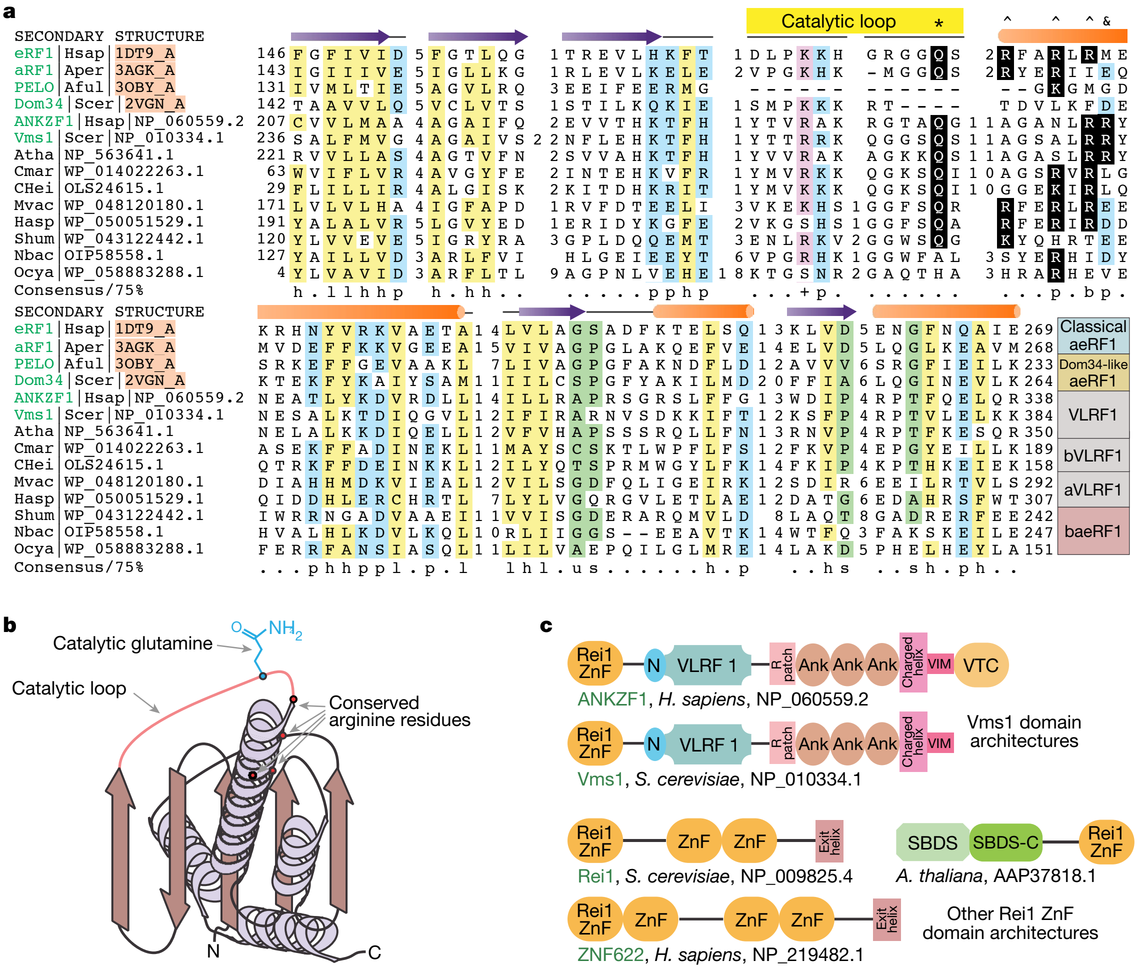


Fig. 2 | Vms1 is a member of the aeRF1 superfamily and the founding member of the VLRF1 clade. a, Multiple sequence alignment of the aeRF1 superfamily; protein names, organism abbreviations and accession numbers are indicated on the left, and families are indicated in boxes on the bottom right. Aful, *Archaeoglobus fulgidus*; Aper, *Aeropyrum pernix*; Atha, *Arabidopsis thaliana*; aVLRF1, archaeal VLRF1 family; baeRF1, bacterial aeRF1 clade; bVLFR1, bacteroidetes VLRF1 family; Cmar, *Cyclobacterium marinum*; CHei, *Candidatus Heimdallarchaeota archaeon*; Hsap, *Homo sapiens*; Hasp, *Halostagnicola* sp. A56; Mvac, *Methanosarcina vacuolata*; Nbac, Nitrospirae bacterium; Ocya, Oscillatoriales cyanobacterium; Scer, *Saccharomyces cerevisiae*; Shum, *Sinomonas humi*. Residues are coloured per consensus conservation, residue characteristics are indicated on the bottom consensus line (h, hydrophobic; l, aliphatic; p, polar; +, positively charged; b, big; u, tiny; s, small), key residues and features are marked at the top ('*' indicates the catalytic glutamine, '^' indicate arginine residues conserved across multiple families and '&' indicates the arginine that is conserved in the VLRF1 family). **b**, Topology diagram of the aeRF1 RNase H fold domain with key features labelled. **c**, Domain architectures of proteins containing the Rei1 C2H2-type zinc-finger. Ank, ankyrin repeat; ZnF, zinc-finger domain.

Fig. 4a). For the substrate, we used 60S fractions from *vms1*Δ cells expressing PrANS. Recombinant Vms1 and the substrate were mixed, and after incubation, peptidyl-tRNA was precipitated using hexadecyl-trimethyl-ammonium bromide (CTAB). His6–Vms1 promoted the release of both Ub–PrANS–tRNA and PrANS–tRNA species from 60S subunits (Fig. 4a). By contrast, His6–Vms1-Q295L was unable to release modified substrate. This assay was reproducible but required extensive manipulations that resulted in spontaneous hydrolysis of the substrate during its preparation. In addition, it was difficult to reliably quantify heterogeneous ubiquitin conjugates by immunoblot. Therefore, we turned to in vitro translation in rabbit reticulocyte lysate, which has been used to reconstitute ubiquitylation of RQC substrates[9, 21]. These earlier studies demonstrated that ubiquitylation of the nascent chain is delayed by approximately 15 min after translation. Because the Vms1-dependent reaction (Fig. 4a) did not display strong dependence on substrate ubiquitylation, we ran brief translation reactions with a non-stop template encoding Flag-tagged CRP with a 30-nucleotide poly(A) tail (Flag–CRPNSKn) to generate stalled nascent chains that had not yet been ubiquitylated, which simplified quantification. The major translation product (Fig. 4b) was confirmed to be tRNA-conjugated Flag–CRPNSKn using CTAB precipitation and RNase treatment (Extended Data Fig. 4b). Consistent with a previous report[9], the stalled Flag–CRPNSKn–tRNA was associated maximally with 80S fractions (Extended Data Fig. 4c).

To assay for eRF1-like peptidyl-tRNA hydrolytic activity, wild-type or mutant His6–Vms1 was incubated with ^{35}S-labelled substrate for increasing periods of time (Fig. 4b). De-acylation of Flag–CRPNSKn–tRNA with concomitant formation of unmodified substrate was observed within three minutes (Fig. 4b, c) and at concentrations as low as 25 nM His6–Vms1 (Fig. 4d, quantified in Extended Data Fig. 4d), whereas essentially no hydrolysis above background was observed with His6–Vms1-Q295L. Because recombinant yeast Vms1 was functional in this heterologous system, we next tested ANKZF1, the human Vms1 orthologue. Purified ANKZF1 (Extended Data Fig. 4a) catalysed deacylation, whereas the active site mutant ANKZF1-Q246L was inactive (Fig. 4e and Extended Data Fig. 4e).

Like *vms1*Δ mutants, *cdc48* mutant cells accumulate Ub–PrANS–tRNA on 60S subunits[7]. This raises the question as to what the role of Cdc48 is in the hydrolysis of peptidyl-tRNA. We were unable to establish ATP-dependence for the in vitro deacylation reaction (data not shown), which perhaps is not surprising because the ATPase Rli1 accelerates nascent chain release by eRF1 in the presence of non-hydrolysable ATP[5]. We suggest that Cdc48 facilitates presentation of the catalytic glutamine of Vms1 at the PTC, much as Rli1 positions

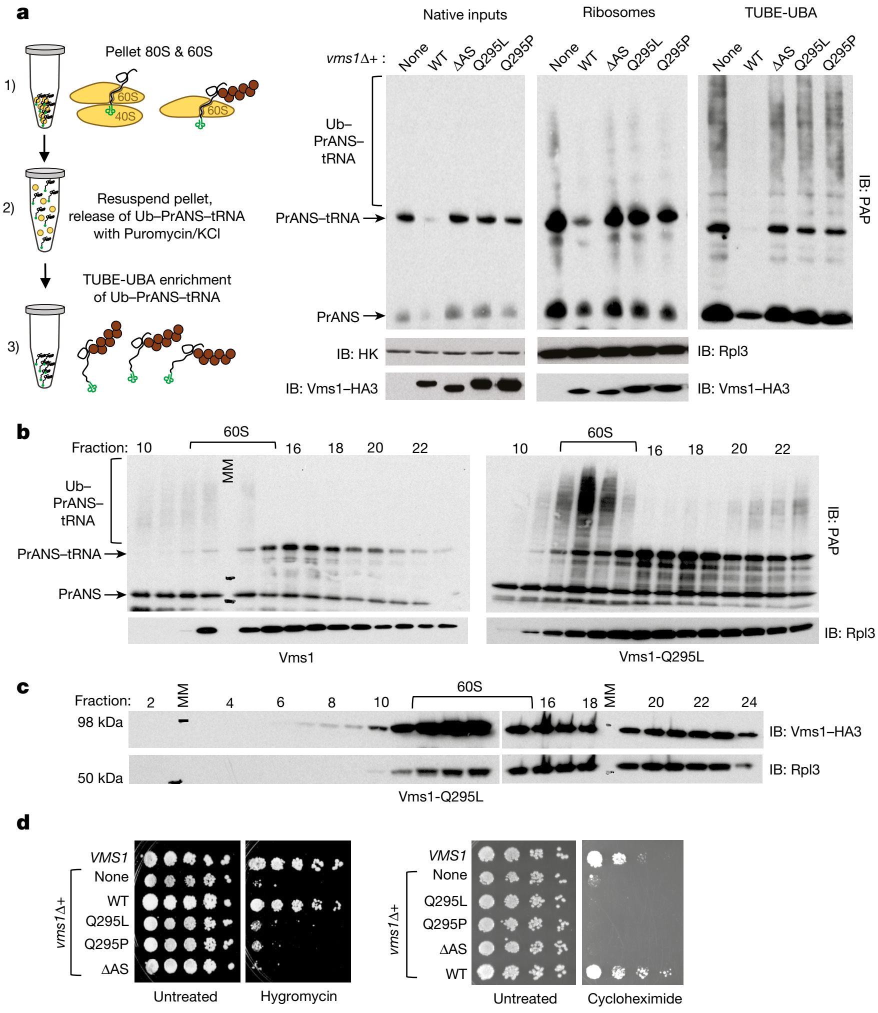


Fig. 3 | Point mutations in the conserved, putative peptidyl-tRNA hydrolase active site of Vms1 phenocopy *vms1*Δ. **a**, Schematic of experimental design. Brown circles, ubiquitin; green circles, tRNA. Left blot, lysates from *vms1*Δ cells expressing either wild-type Vms1–HA3 or the active site mutants were fractionated on a NuPAGE gel, and immunoblotted to detect protein A and Vms1–HA3. Hexokinase (HK) served as a loading control. ΔAS lacks amino acids 283–314. Middle blot, ribosomes were isolated on sucrose cushions, and aliquots were immunoblotted to detect protein A, Rpl3 and Vms1–HA3. Right blot, the remaining ribosomes were adsorbed to TUBE resin and bound fractions were resolved using Tris–glycine gels and immunoblotted to detect protein A. **b**, Native lysates (10 A_{260} units) from *vms1*Δ expressing PrANS and either wild-type Vms1–HA3 (left) or the Q295L mutant (right) were fractionated on sucrose gradients and immunoblotted to detect protein A and Rpl3. Molecular mass markers indicate 36 and 23 kDa. **c**, Sucrose gradient fractions from *vms1*Δ expressing Vms1-Q295L–HA3 were immunoblotted to detect mutant Vms1 and Rpl3. They were run at the same time as wild-type Vms1–HA3 in Fig. 1e and the exposures are identical. All western blots are representative of two biological replicates. Gel source images are provided in Supplementary Fig. 1. **d**, Serial tenfold dilutions of wild-type and *vms1*Δ cells transformed with the indicated *vms1* alleles were spotted on YPD plates containing 100 μg ml^{-1} hygromycin (left) or 25 ng ml^{-1} cycloheximide (right) and incubated at 30 °C for 3 days. Images shown are representative of three biological replicates.

eRF1 in canonical termination[28]. A scaffolding function could account for accumulation of ubiquitylated reporter in 60S and 80S fractions in *vms1*$^{\Delta VIM}$ mutants (Extended Data Fig. 3c). Cdc48 ATPase working together with Ltn1 may have an additional role at the exit tunnel after deacylation. The emerging nascent chain is ubiquitylated by Ltn1, which recruits Cdc48–Ufd1–Npl4[6, 7]. The latter can then translocate along and unfold the nascent chain[15, 16] thereby ensuring its extraction from the 60S subunit. In other contexts, different factors may facilitate nascent chain clearance. For example, defective ribosome–nascent chain complexes targeted to mitochondria become stuck during import[29]. Vms1 promotes clearance of these stalls in a VIM-independent manner. Owing to close apposition of the stalled ribosome–nascent chain and mitochondrial translocon[29], neither Ltn1 nor Cdc48–Ufd1–Npl4 should gain access to the deacylated nascent chain to ensure its extraction. Vms1 also localizes to the endoplasmic reticulum and the RQC pathway can rescue nascent chains that are stalled at the Sec61 translocon[30].

Three observations point to an additional role for Vms1 in the release of stalled nascent chain–tRNA from intact ribosomes, which is also likely to be Ltn1-independent: (i) Vms1 associates with all forms of ribosomes and this persists in *ltn1*Δ (Fig. 1e), (ii) *vms1* mutant cells exhibited strong accumulation of PrANS–tRNA in both monosome and polysome fractions (Figs. 1d and 3b), and (iii) Low doses (Fig. 4d) of Vms1 rapidly (Fig. 4b) deacylated peptidyl-tRNA that was associated with the 80S ribosome. Nascent chain-engaged 60S subunits can, however, re-associate with free 40S subunits in vitro to generate 80S ribosome–nascent chain complexes[21].

We suggest that in vivo, the leading stalled ribosome on non-stop poly(A) mRNA that has its A site occupied could be the monosome substrate for Vms1 (see Extended Data Fig. 5 for model). Cleavage by an unidentified endonuclease generates ribosomes with empty A sites on truncated mRNA that are known to be rescued by Dom34–Hbs1–Rli1[1]. Jammed nascent chain–tRNA on these split 60S subunits is ubiquitylated by Ltn1 and the tRNA linkage is hydrolysed by Vms1. Our

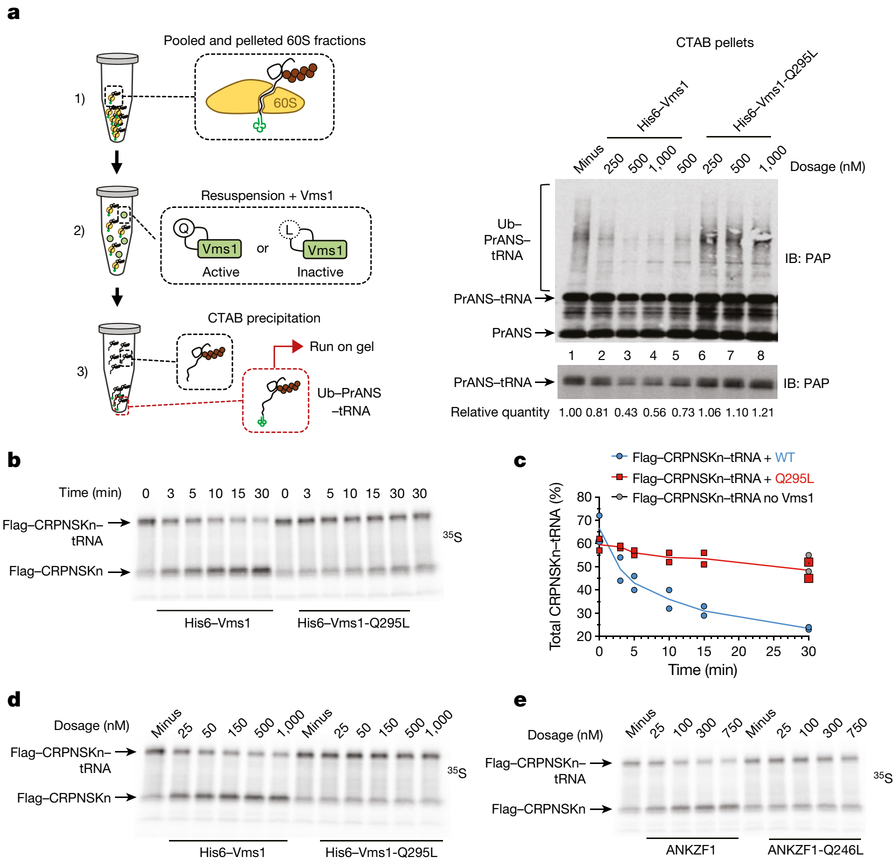


Fig. 4 | Peptidyl-tRNA hydrolase activity of Vms1 and ANKZF1 depend on the catalytic glutamine residue. a, Left, experimental design. Brown circles, ubiquitin; green, tRNA. Lysate from *vms1*Δ cells expressing PrANS was fractionated on a sucrose gradient, and 60S fractions were pooled and pelleted. Top right, His6–Vms1 or His6–Vms1-Q295L was added and incubated for 10 min at 30 °C. Ub–PrANS–tRNA was precipitated with CTAB, and immunoblotted with PAP to detect protein A. Bottom right, shorter exposure from which the discrete peptidyl-tRNA band was quantified (lane five contained additional 1 mM GTP). Western blot is representative of two biological replicates. Minus, no recombinant Vms1 added. **b**, Time course of deacylation reaction using ribosome–nascent chain complexes generated by in vitro translation. Flag–CRPNSKn was translated in rabbit reticulocyte lysate in the presence of ^{35}S-labelled methionine for 30 min at 30 °C. Recombinant wild-type or mutant His6–Vms1 (150 nM) was added to ribosome–nascent chains (75–100 nM) for the indicated time periods. Aliquots were visualized by autoradiography. **c**, Phosphorimager quantification of two independent biological replicates from **b**. WT, His6–Vms1; Q295L, His6–Vms1-Q295L. **d**, Titration of Vms1. Reactions and processing as in **b**. **e**, Wild-type ANKZF1 and ANKZF1-Q246L mutant were analysed as in **b**, except that the hydrolase reaction was performed at 37 °C. The image shown is representative of two biological replicates. Gel source images and quantification data are provided in Supplementary Fig. 1.

discovery of Vms1 as the founding member of the VLRF1 clade reveals that multiple different functional peptidyl-tRNA hydrolase release factors belonging to the aeRF1 superfamily coexist in eukaryotic cells. We posit that VLRF1 and Dom34 evolved to protect the translation apparatus from errors in transcription and post-transcriptional processing that generate defective mRNAs.

Online content

Any Methods, including any statements of data availability and Nature Research reporting summaries, along with any additional references and Source Data files, are available in the online version of the paper at https://doi.org/10.1038/s41586-018-0022-5.

Received: 21 August 2017; Accepted: 8 February 2018;
Published online 9 April 2018.

1. Buskirk, A. R. & Green, R. Ribosome pausing, arrest and rescue in bacteria and eukaryotes. *Phil. Trans. R. Soc. Lond. B* **372**, 20160183 (2017).
2. Joazeiro, C. A. P. Ribosomal stalling during translation: providing substrates for ribosome-associated protein quality control. *Annu. Rev. Cell Dev. Biol.* **33**, 343–368 (2017).
3. Ito-Harashima, S., Kuroha, K., Tatematsu, T. & Inada, T. Translation of the poly(A) tail plays crucial roles in nonstop mRNA surveillance via translation repression and protein destabilization by proteasome in yeast. *Genes Dev.* **21**, 519–524 (2007).
4. Bengtson, M. H. & Joazeiro, C. A. Role of a ribosome-associated E3 ubiquitin ligase in protein quality control. *Nature* **467**, 470–473 (2010).
5. Shoemaker, C. J. & Green, R. Kinetic analysis reveals the ordered coupling of translation termination and ribosome recycling in yeast. *Proc. Natl Acad. Sci. USA* **108**, E1392–E1398 (2011).
6. Brandman, O. et al. A ribosome-bound quality control complex triggers degradation of nascent peptides and signals translation stress. *Cell* **151**, 1042–1054 (2012).
7. Verma, R., Oania, R. S., Kolawa, N. J. & Deshaies, R. J. Cdc48/p97 promotes degradation of aberrant nascent polypeptides bound to the ribosome. *eLife* **2**, e00308 (2013).
8. Defenouillère, Q. et al. Cdc48-associated complex bound to 60S particles is required for the clearance of aberrant translation products. *Proc. Natl Acad. Sci. USA* **110**, 5046–5051 (2013).
9. Shao, S., von der Malsburg, K. & Hegde, R. S. Listerin-dependent nascent protein ubiquitination relies on ribosome subunit dissociation. *Mol. Cell* **50**, 637–648 (2013).
10. Shen, P. S. et al. Protein synthesis. Rqc2p and 60S ribosomal subunits mediate mRNA-independent elongation of nascent chains. *Science* **347**, 75–78 (2015).
11. Chu, J. et al. A mouse forward genetics screen identifies LISTERIN as an E3 ubiquitin ligase involved in neurodegeneration. *Proc. Natl Acad. Sci. USA* **106**, 2097–2103 (2009).
12. Choe, Y. J. et al. Failure of RQC machinery causes protein aggregation and proteotoxic stress. *Nature* **531**, 191–195 (2016).
13. Yonashiro, R. et al. The Rqc2/Tae2 subunit of the ribosome-associated quality control (RQC) complex marks ribosome-stalled nascent polypeptide chains for aggregation. *eLife* **5**, e11794 (2016).
14. Defenouillère, Q. et al. Rqc1 and Ltn1 prevent C-terminal alanine-threonine tail (CAT-tail)-induced protein aggregation by efficient recruitment of Cdc48 on stalled 60S subunits. *J. Biol. Chem.* **291**, 12245–12253 (2016).
15. Blythe, E. E., Olson, K. C., Chau, V. & Deshaies, R. J. Ubiquitin- and ATP-dependent unfoldase activity of P97/VCP•NPLOC4•UFD1L is enhanced by a

mutation that causes multisystem proteinopathy. *Proc. Natl Acad. Sci. USA* **114**, E4380–E4388 (2017).
16. Bodnar, N. O. & Rapoport, T. A. Molecular mechanism of substrate processing by the Cdc48 ATPase complex. *Cell* **169**, 722–735.e9 (2017).
17. Hänzelmann, P. & Schindelin, H. The structural and functional basis of the p97/valosin-containing protein (VCP)-interacting motif (VIM): mutually exclusive binding of cofactors to the N-terminal domain of p97. *J. Biol. Chem.* **286**, 38679–38690 (2011).
18. Stapf, C., Cartwright, E., Bycroft, M., Hofmann, K. & Buchberger, A. The general definition of the p97/valosin-containing protein (VCP)-interacting motif (VIM) delineates a new family of p97 cofactors. *J. Biol. Chem.* **286**, 38670–38678 (2011).
19. van Hoof, A., Frischmeyer, P. A., Dietz, H. C. & Parker, R. Exosome-mediated recognition and degradation of mRNAs lacking a termination codon. *Science* **295**, 2262–2264 (2002).
20. Alamgir, M., Erukova, V., Jessulat, M., Azizi, A. & Golshani, A. Chemical-genetic profile analysis of five inhibitory compounds in yeast. *BMC Chem. Biol.* **10**, 6 (2010).
21. Shao, S. & Hegde, R. S. Reconstitution of a minimal ribosome-associated ubiquitination pathway with purified factors. *Mol. Cell* **55**, 880–890 (2014).
22. Nielson, J. R. et al. Sterol oxidation mediates stress-responsive Vms1 translocation to mitochondria. *Mol. Cell* **68**, 673–685.e6 (2017).
23. Song, H. et al. The crystal structure of human eukaryotic release factor eRF1—mechanism of stop codon recognition and peptidyl-tRNA hydrolysis. *Cell* **100**, 311–321 (2000).
24. Jin, H., Kelley, A. C., Loakes, D. & Ramakrishnan, V. Structure of the 70S ribosome bound to release factor 2 and a substrate analog provides insights into catalysis of peptide release. *Proc. Natl Acad. Sci. USA* **107**, 8593–8598 (2010).
25. Cheng, Z. et al. Structural insights into eRF3 and stop codon recognition by eRF1. *Genes Dev.* **23**, 1106–1118 (2009).
26. Korostelev, A. et al. Crystal structure of a translation termination complex formed with release factor RF2. *Proc. Natl Acad. Sci. USA* **105**, 19684–19689 (2008).
27. Greber, B. J. et al. Insertion of the biogenesis factor Rei1 probes the ribosomal tunnel during 60S maturation. *Cell* **164**, 91–102 (2016).
28. Preis, A. et al. Cryoelectron microscopic structures of eukaryotic translation termination complexes containing eRF1-eRF3 or eRF1-ABCE1. *Cell Rep.* **8**, 59–65 (2014).
29. Izawa, T., Park, S. H., Zhao, L., Hartl, F. U. & Neupert, W. Cytosolic protein Vms1 links ribosome quality control to mitochondrial and cellular homeostasis. *Cell* **171**, 890–903.e18 (2017).
30. von der Malsburg, K., Shao, S. & Hegde, R. S. The ribosome quality control pathway can access nascent polypeptides stalled at the Sec61 translocon. *Mol. Biol. Cell* **26**, 2168–2180 (2015).

Acknowledgements We thank A. Buchberger, H. Rao, A. van Hoof, R. Voorhees and J. Warner for reagents; the Proteome Exploration Laboratory (PEL), Caltech, for help with the mass spectrometry analysis and the members of the Deshaies laboratory, M. Blanco, M. Guttman, B. Clemons, R. Voorhees and S. Shan for discussions. R.J.D. was an investigator of the HHMI and this work was funded in part by the HHMI. A.M.B. and L.A. are supported by the funds of the Intramural Research Program of the National Library of Medicine.

Author contributions R.V., K.M.R. and R.J.D. designed experiments, J.M.R. performed mass spectrometry analyses, L.A. and A.M.B. did the computational analyses, K.M.R. performed the mutagenesis analyses, R.S.O. generated all the yeast strains and R.V. performed all the biochemical experiments. R.V. and R.J.D. supervised research, R.V., R.J.D. and L.A. wrote the paper and all authors participated in editing the manuscript. All figure schematics were generated by K.M.R.

Competing interests R.V., K.M.R., A.M.B., R.S.O., J.M.R. and L.A. declare no competing interests. R.J.D. is currently Senior Vice President of discovery research at Amgen and a Visiting Associate at the California Institute of Technology (Caltech).

Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0022-5.
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0022-5.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to L.A. or R.J.D.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

Yeast strains and culture conditions. All strains used in the current study are listed in Extended Data Table 1. They were derived from W303 or S288C backgrounds and grown as described[7].

Western blot analyses. The following antibodies were used: PAP (Sigma, P1291), anti-Flag M2 (Sigma, F-1804), anti-TAP (Thermo Fisher, CAB1001) and anti-HA 3F10 (Roche/Sigma, 12013819001). Anti-Rpl32 and -Rpl3 antibodies were a gift from J. Warner.

Ribosome isolation and sub-fractionation. The sucrose cushion method for the isolation of ribosomes; the binding of ribosomal preparations to TUBE-UBA resin; sucrose density gradient centrifugation; and CTAB precipitation of peptidyl-tRNA from sucrose gradient fractions were performed as previously described[7]. TUBE resin was purchased from Boston Biochem (AM-130). CTAB was purchased from Sigma (52365).

Site-directed mutagenesis. All expression constructs for mutational analysis were generated using the Q5-Site-Directed Mutagenesis Kit (NEB E00554S) following the manufacturer's instructions. Primers were designed using the online tool NEBaseChanger (http://nebasechanger.neb.com). Mutations were confirmed by Sanger sequencing. All plasmids are listed in Extended Data Table 1.

Purification of His6–Vms1 and His6–Vms1-Q295L from bacteria. Wild-type and mutant Vms1 (Extended Data Table 1) were cloned into the bacterial pET-28a vector in frame with the N-terminal His6 tag. Expression was induced in Rosetta cells using 1 mM IPTG at 16 °C for 16 h. Cell pellets were brought up in lysis buffer containing 50 mM Hepes, pH 8.0, 0.5 M NaCl, 20 mM imidazole and 0.5% Triton, sonicated, and centrifuged at 34,000*g* for 20 min in a Sorvall SS34 rotor. Supernatant was bound to Ni–NTA affinity resin (Qiagen) for 1 h at 4 °C. Bound proteins were washed four times in lysis buffer and eluted in 25 mM Hepes, pH 7.5, 150 mM NaCl, 5% glycerol and 250 mM imidazole. Imidazole was removed by dialysis and proteins were concentrated using Amicon Ultra centrifugal filters.

Purification of ANKZF1–Flag–Myc and ANKZF1(Q246L)–Flag–Myc. Plasmids were constructed that contained the cytomegalovirus (CMV) promoter followed by sequences encoding wild-type (Origene) or Q246L ANKZF1 (current study; confirmed by sequencing) tagged with a Flag–Myc epitope sequence at the 3′ end of the open reading frame. These constructs were transiently transfected into human embryonic kidney HEK293 cells (ATCC, tested negative for mycoplasma contamination). After 72 h, proteins were purified using a previously published protocol for Flag-tagged mammalian proteins[21].

In vitro transcription and translation. PCR products were generated using 3×Flag–CRP (Extended Data Table 1) as a template with the reverse primer either containing a stop codon (3×Flag–CRPStop), or lacking a stop codon but with a 30-nucleotide poly(T) extension (3×Flag-CRPNSKn). PCR products were transcribed and translated in vitro using TnT T7 Quick rabbit reticulocyte lysate (Promega, L5540) in the presence of radioactive methionine (Perkin Elmer, NEG709A). Reactions were typically for 30 min at 30 °C, at which time recombinant proteins were added for another 10 min. Aliquots were resolved by SDS–PAGE, and dried gels were exposed either to film or to a PhosphorImager screen.

Isolation of tandem affinity purification-tagged complexes from yeast. Cells in which *CDC48*, *VMS1* or *UFD1* were tagged at the 3′ end with the tandem affinity purification (TAP) epitope sequence[31] were grown to exponential phase in 2 l YPD medium and collected. Washed cell pellets were ground in liquid nitrogen, and the powder was resuspended in a lysis buffer containing 25 mM Tris, pH 7.5, 150 mM NaCl, 0.25% Triton and protease inhibitor tablets (Roche, 11873580001) at a ratio of two volumes of powder to one volume of buffer. Following centrifugation at 34,000*g* for 20 min (Sorvall SS34 rotor), lysates were bound to 200 μl acid-washed IgG sepharose 6 Fast Flow (GE Healthcare, 17-0969-01) at 4 °C for 90 min. Bound proteins were washed twice in lysis buffer, twice in TEV buffer (25 mM Tris, pH 7.5, 150 mM NaCl, 2 mM dithiothreitol, 5% glycerol) and eluted by cleavage with 50 U TEV for 2 h at 16 °C.

Mass spectrometry analyses. Eluted samples were lyophilized and digested for mass spectrometry analysis as described[32]. Approximately 250 ng of trypsin-digested peptides were loaded onto a 26-cm analytical HPLC column (75-μm inner diameter) packed in-house with ReproSil-Pur C_{18AQ} 1.9-μm resin (120 Å pore size; Dr. Maisch). Once loaded, the peptides were separated with a 120-min gradient at a flow rate of 350 nl min^{-1} at 50 °C (column heater) using the following gradient: 2–6% solvent B (7.5 min), 6–25% B (82.5 min), 25–40% B (30 min), 40–100% B (1 min) and 100% B (9 min), in which solvent A was 97.8% H_2O, 2% acetonitrile and 0.2% formic acid, and solvent B was 19.8% H_2O, 80% acetonitrile and 0.2% formic acid. Samples were analysed using an EASY-nLC 1000 coupled to an Orbitrap Fusion operated in data-dependent acquisition mode to automatically switch between a full scan ($m/z = 350$–1,500) in the Orbitrap at 120,000 resolving power and a tandem mass spectrometry scan of higher-energy collisional dissociation fragmentation that was detected in the ion trap (using TopSpeed). The automatic gain control targets of the Orbitrap and ion trap were 400,000 and 10,000, respectively.

Mass spectrometry data analysis. All raw data were searched concurrently using MaxQuant (v.1.5.3.30[33,34]) against the *Saccharomyces* Genome Database (SGD) and a common contaminant database. Precursor mass tolerance was 4.5 p.p.m. after automatic recalibration. Fragment ion tolerance was 0.5 Da. Searches permitted up to two missed tryptic peptide cleavages. Cysteine carbamidomethylation was designated as a fixed modification while methionine oxidation and N-terminal acetylation were designated as variable modifications. Intensity-based absolute quantification (iBAQ) protein quantification and 'match between runs' were enabled. Protein and peptide false discovery rates were estimated to be less than 1% using a target-decoy approach.

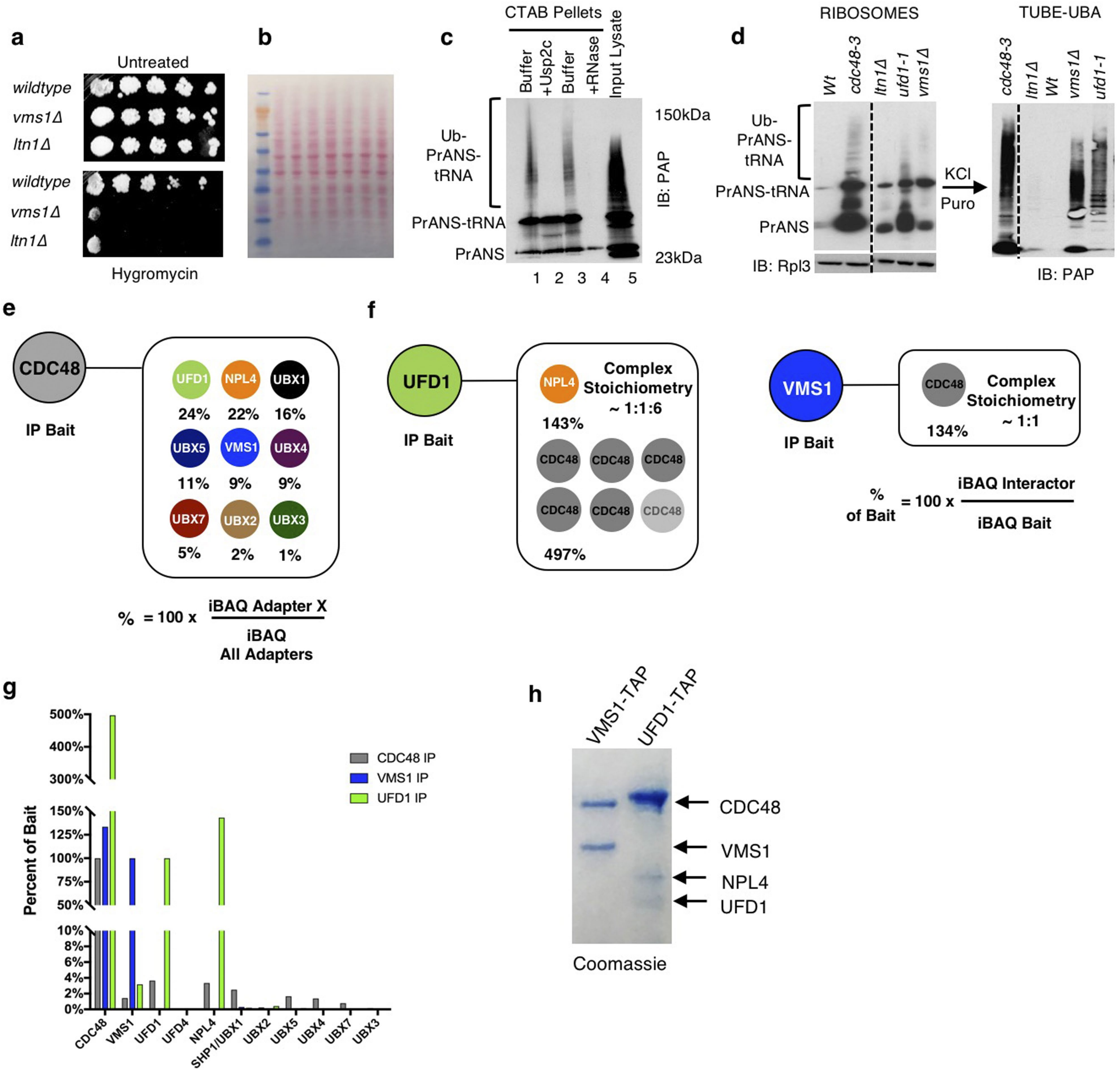
Computational analyses. Iterative sequence profile searches were performed using the PSI-BLAST[35] and JACKHMMER[36] programs run against the NCBI non-redundant protein database clustered at 50% identity using the MMseqs2 program. Similarity-based clustering for both classification and culling of nearly identical sequences was performed using the BLASTCLUST program (ftp://ftp.ncbi.nih.gov/blast/documents/blastclust.html). The length (*L*) and score (*S*) threshold parameters were variably adjusted depending on need. For example, the length and score threshold parameters for clustering near identical proteins were $L = 0.9$ and $S = 1.2$. The HHpred program was used for profile–profile searches[37]. Structure similarity searches were performed using the DaliLite program[38]. Multiple sequence alignments were made using the Kalign program followed by manual adjustments on the basis of profile–profile searches and structural alignments[39]. Secondary structures were predicted using the JPred program[40]. For previously known domains, the Pfam database[41] was used as a guide, although the profiles were augmented by addition of newly detected divergent members that were not detected by the original Pfam models. Clustering with BLASTCLUST combined with multiple sequence alignment and further sequence profile searches were used to identify other domains that were not present in the Pfam database. Contextual information from prokaryotic gene neighbourhoods was retrieved by a Perl script that extracts the upstream and downstream genes of the query gene and uses BLASTCLUST to cluster the proteins to identify conserved gene-neighbourhoods. Phylogenetic analysis was conducted using an approximately-maximum-likelihood method implemented in the FastTree 2.1 program under default parameters[42]. Structural visualization and manipulations were performed using the PyMol v.1.8.2.0 (https://www.pymol.org) program.

Data reporting. All data shown are representative of at least two independent biological experiments.

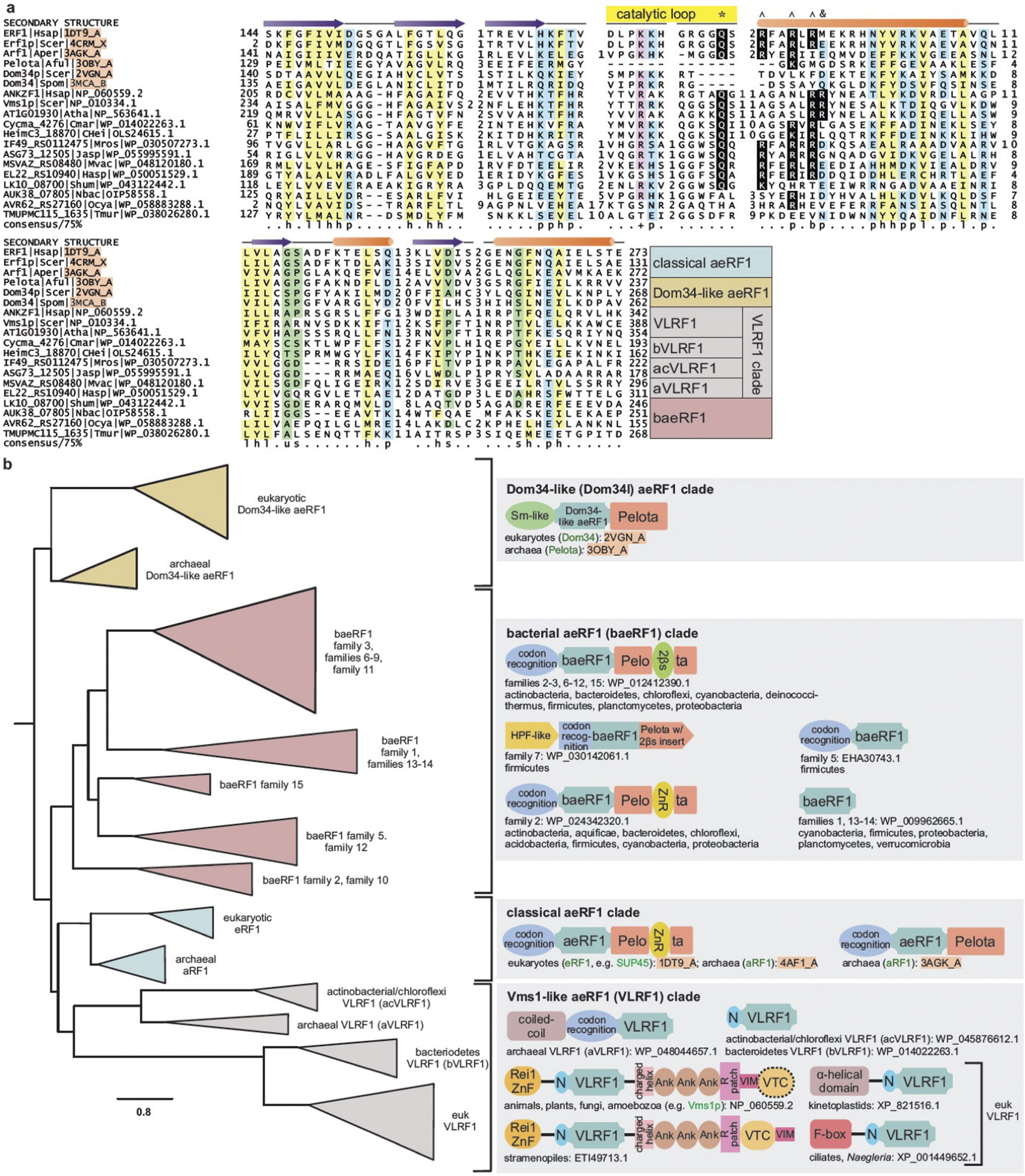
Reporting summary. Further information on experimental design is available in the Nature Research Reporting Summary linked to this paper.

Data availability. Gel source images for Figs. 1, 3, 4 and Extended Data Figs. 1, 3 and 4 are available in Supplementary Fig. 1. Source Data for quantifications in Fig. 4 and Extended Data Fig. 4 are provided in the online version of the paper. All other data supporting the findings of this study are available from the corresponding authors upon reasonable request.

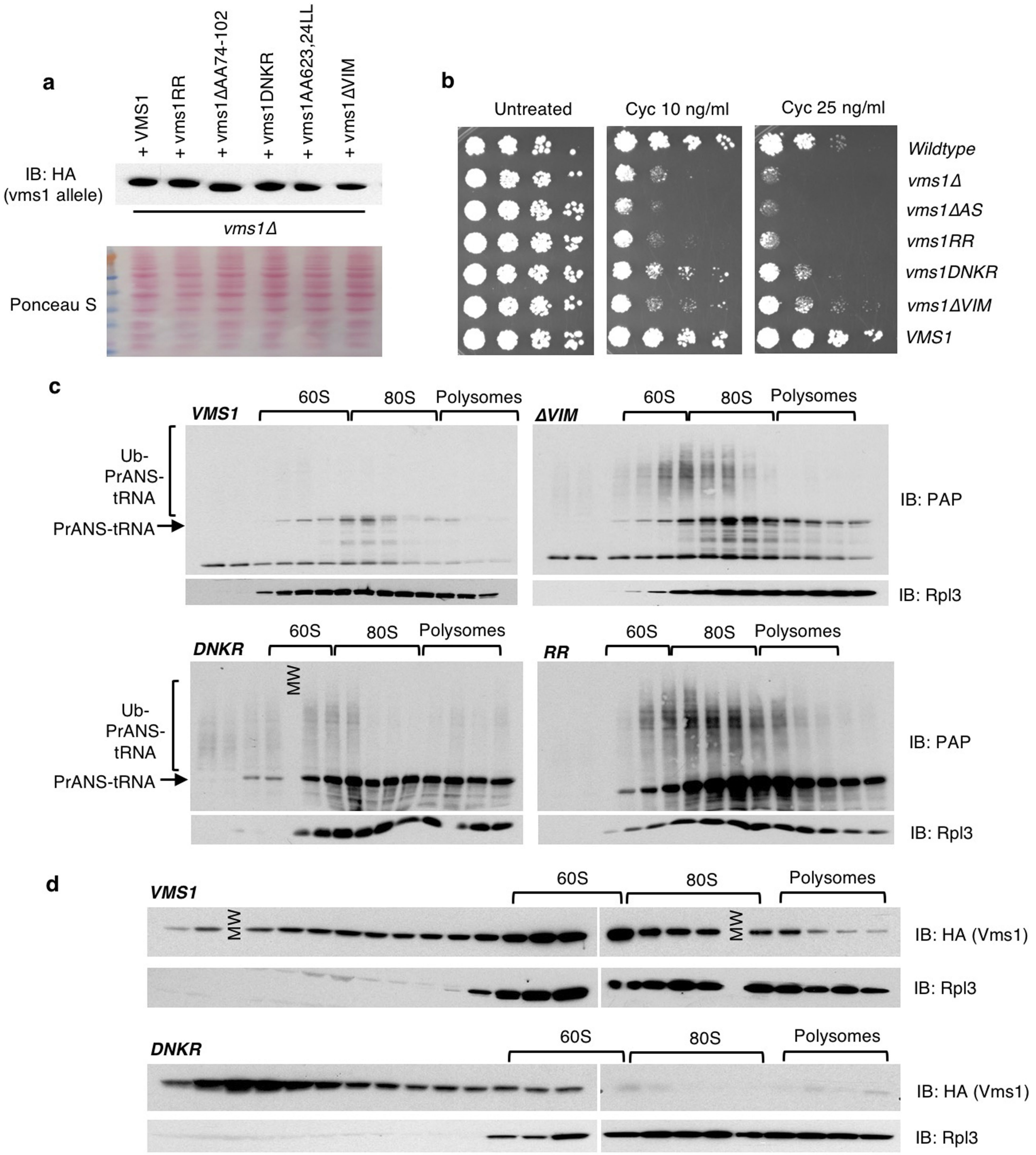
31. Puig, O. et al. The tandem affinity purification (TAP) method: a general procedure of protein complex purification. *Methods* **24**, 218–229 (2001).
32. Pierce, N. W. et al. Cand1 promotes assembly of new SCF complexes through dynamic exchange of F box proteins. *Cell* **153**, 206–215 (2013).
33. Cox, J. & Mann, M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol.* **26**, 1367–1372 (2008).
34. Wagner, S. A. et al. A proteome-wide, quantitative survey of in vivo ubiquitylation sites reveals widespread regulatory roles. *Mol. Cell Proteomics* **10**, M111.013284 (2011).
35. Altschul, S. F. et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402 (1997).
36. Eddy, S. R. A new generation of homology search tools based on probabilistic inference. *Genome Inform.* **23**, 205–211 (2009).
37. Alva, V., Nam, S. Z., Söding, J. & Lupas, A. N. The MPI bioinformatics Toolkit as an integrative platform for advanced protein sequence and structure analysis. *Nucleic Acids Res.* **44**, W410–W415 (2016).
38. Holm, L., Kääriäinen, S., Rosenström, P. & Schenkel, A. Searching protein structure databases with DaliLite v.3. *Bioinformatics* **24**, 2780–2781 (2008).
39. Lassmann, T., Frings, O. & Sonnhammer, E. L. Kalign2: high-performance multiple alignment of protein and nucleotide sequences allowing external features. *Nucleic Acids Res.* **37**, 858–865 (2009).
40. Cole, C., Barber, J. D. & Barton, G. J. The Jpred 3 secondary structure prediction server. *Nucleic Acids Res.* **36**, W197–W201 (2008).
41. Finn, R. D. et al. The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res.* **44**, D279–D285 (2016).
42. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2—approximately maximum-likelihood trees for large alignments. *PLoS ONE* **5**, e9490 (2010).



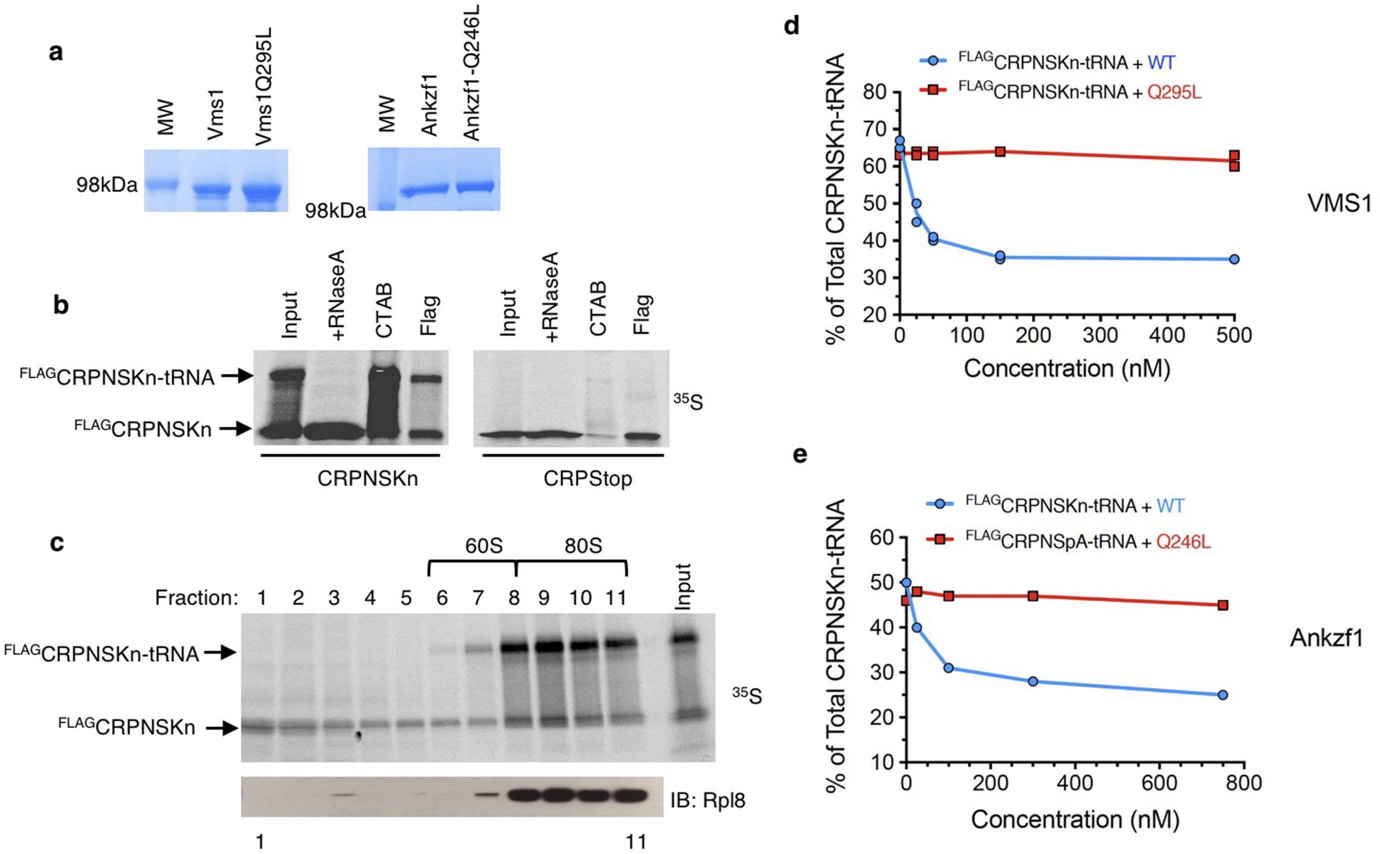
Extended Data Fig. 1 | Mass spectrometry analysis of Vms1 required for release of Ub–PrANS–tRNA from ribosomes. a, Serial tenfold dilutions of exponential cultures were spotted on YPD plates with or without 100 μg ml^{-1} hygromycin B and allowed to grow at 30 °C for 3 days. Data shown are representative of three biological replicates. **b**, Ponceau S stained nitrocellulose filter used as loading control for Fig. 1c. **c**, Lysate from *vms1Δ* cells expressing PrANS (input lysate, lane five) was fractionated on a sucrose gradient and 60S fractions were pooled and mock-treated (lanes one and three) or pretreated with the deubiquitylating enzyme Usp2c (1 μM; lane two) or 100 μg ml^{-1} RNase A (lane four) at 30 °C for 20 min before CTAB precipitation. The pellets were immunoblotted with PAP. **d**, Ribosomes from the indicated strains were isolated using sucrose cushions and aliquots were immunoblotted to detect PrANS and Rpl3. The remainder was bound to TUBE resin. The adsorbed fractions were immunoblotted with PAP. All lanes in the left and right panels are from the same blots. The dashed lines indicate cropping of lanes not pertinent to the current study. Uncropped gel source images are provided in Supplementary Fig. 1. Data in **c** and **d** are representative of two biological replicates. **e–h**, Mass spectrometry analysis of TAP-tagged Cdc48, Vms1 and Ufd1. **e**, Relative abundance of each Cdc48 adaptor co-immunoprecipitated with Cdc48 relative to all adaptor proteins identified. **f**, Schematic illustrating the estimated stoichiometry of Cdc48–adaptor complexes. **g**, Relative stoichiometry of associated proteins that co-immunoprecipitated with Cdc48, Vms1 and Ufd1. Samples were normalized to the iBAQ value of the bait protein and are presented as percentage of the bait protein. Protein iBAQ values from the untagged control were subtracted from the tagged immunoprecipitation samples. **h**, Coomassie blue-stained gel of samples used for mass spectrometry analysis. All mass spectrometry data (**e–h**) are representative of two biological replicates.



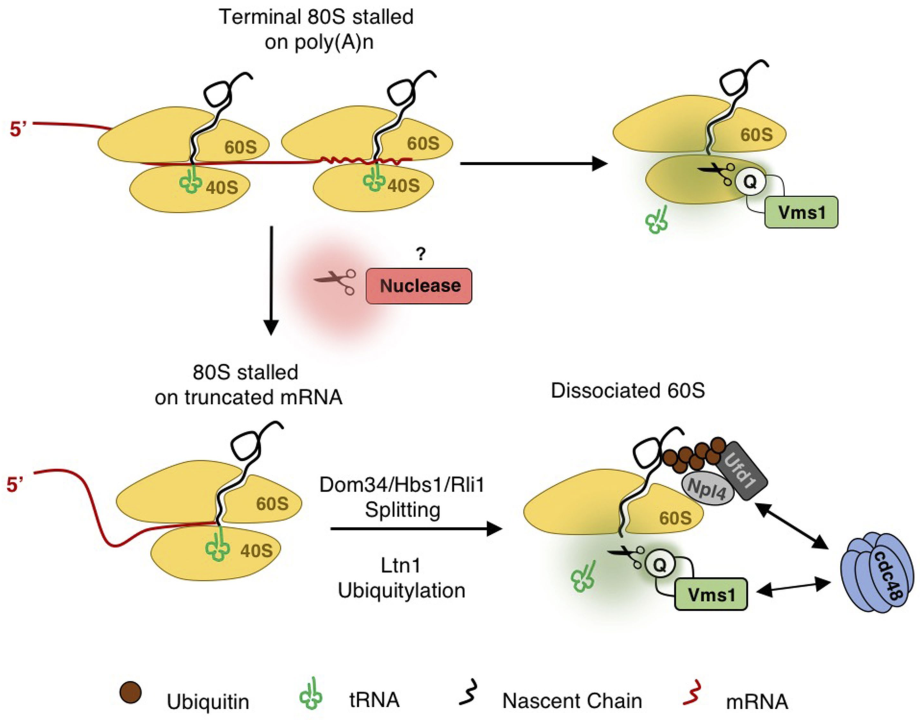
Extended Data Fig. 2 | Vms1 is the founding member of the VLRF1 clade. a, Extended sequence alignment of aeRF1 superfamily with representatives from all families and clades (compared to the limited subset shown in Fig. 2a). **b**, Phylogenetic tree depicting relationships within the aeRF1 superfamily; colouring matches clade labels in Fig. 2a. In the classical aeRF1 clade, two branches contain eukaryotic orthologues (eRF1) and archaeal orthologues (aRF1), respectively. Of the bacterial (baeRF1) versions, certain members are misannotated as 'Host_attach' in the Pfam database while most cannot be detected by existing profiles. The total number of prokaryotic representatives of the VLRF1 clade in the non-redundant database (NCBI, as of 1 December 2017) is 1,044. Of these, the archaeal VLRF1 family (aVLRF1) has 279 members, actino-chloroflexi VLRF1 family (acVLRF1) has 669 and the bacteroidetes VLRF1 family (bVLRF1) has 96. Notable domain architectures and conserved gene neighbourhoods are shown to the right of tree. A gene encoding a ribosome hibernation factor (HPF-like or YfiA-like) that facilitates inactive ribosome aggregation frequently co-occurs and is predicted to function with the baeRF1 domains. The labels below the architecture diagrams list the well-characterized clade representatives, PDB structure accessions and phyletic distributions. The dashed outline indicates domains that are not universally present.



Extended Data Fig. 3 | Mutagenesis of the non-catalytic domains of Vms1. **a**, SDS lysates of cells were immunoblotted with anti-haemagglutinin antibodies. Ponceau S staining of the blot shows the equivalency of the extracts. The RR mutant is R313A, R314A. The DNKR mutant contains the following four mutations in the zinc-finger domain: D94A, N99A, K101A, and R102A. The ΔVIM mutant is deleted for amino acids 622–625. **b**, Serial tenfold dilutions of wild-type and *vms1*Δ cells transformed with the indicated *vms1* alleles were spotted on YPD plates containing 10 or 25 ng ml^{-1} cycloheximide and incubated at 30 °C for 3 days. Data are representative of three biological replicates. **c**, Native lysates (10 A_{260} units each) were subjected to sucrose gradient analysis. In each case fractions 10–23 were resolved and immunoblotted to detect protein A and Rpl3. Identical exposures are shown for all panels. **d**, Native lysates of *vms1*Δ cells expressing wild-type Vms1–HA3 or Vms1–HA3-DNKR were subjected to sucrose gradient analysis. Fractions were immunoblotted for Vms1 and Rpl3. MW, molecular weight marker. All western blot data are representative of two biological replicates. Gel source images are available in Supplementary Fig. 1.



Extended Data Fig. 4 | In vitro reconstitution of Vms1 peptidyl-tRNA hydrolase activity. a, Coomassie blue-stained gels of the indicated purified proteins used for in vitro reconstitution. **b**, Analysis of ribosome–nascent chain complexes generated by translation in reticulocyte lysate. PCR products encoding Flag–CRPNS followed by a 30-nucleotide poly(A) sequence and Flag–CRPStop (with a stop codon) were transcribed and translated in reticulocyte lysate in the presence of ^{35}S-labelled methionine. Completed translation reactions were treated (or not treated), fractionated using SDS–PAGE, and visualized by autoradiography. Lane one, no treatment; lane two, RNase A treatment; lane three, CTAB treatment; lane four, immunoprecipitated with anti-Flag resin. For lanes three and four, the pellet fraction was analysed. **c**, Sucrose gradient analysis of reticulocyte translation reactions. Flag–CRPNSKn was transcribed and translated in 200 μl reticulocyte lysate for 30 min. Lysates were layered onto 2-ml 10–50% sucrose gradients and centrifuged using a Beckman SW55 rotor for 80 min. Eleven fractions (200 μl each) were collected from the top by hand. Aliquots were analysed for fractionation of ^{35}S-labelled substrate and Rpl8 by autoradiography and immunoblotting, respectively. **d**, Quantification of two independent biological replicates of the yeast His6–Vms1 and His6–Vms1-Q295L titration reactions shown in Fig. 4d. **e**, Quantification of the human ANKZF1 and ANKZF1-Q246L titration reactions shown in Fig. 4e. Data are representative of two biological replicates. Gel source images are provided in Supplementary Fig. 1 and Source Data are available with the online version of this paper.



Extended Data Fig. 5 | Working model of Vms1 function at stalled ribosomes. In non-stop decay of mRNAs lacking a stop codon, ribosomes translate the poly(A) tail and stall after translating several lysines[4, 19]. The A site as a consequence is occupied by an AAA codon. Our data suggest that Vms1 can potentially hydrolyse peptidyl-tRNA chains on this leading stalled ribosome without prior splitting by Dom34–Hbs1–Rli1. One of the known responses to stalling is endonucleolytic cleavage of the mRNA by an as-yet unidentified endonuclease. The cleavage reaction generates a truncated transcript. Lagging ribosomes that translate up to the cleavage site stall with an empty A site. Such stalls are recognized by Dom34–Hbs1 that together with Rli1 dissociate the 80S ribosome into the 40S subunit and the 60S subunit, which contains the nascent peptidyl-tRNA[1]. Dissociation allows for stable association of the RQC complex members Rqc1–Rqc2 and the E3 ubiquitin ligase Ltn1[9]. Rqc2 adds non-templated alanine and threonine residues to the C-terminal end of the nascent chain to extrude sequences in the exit tunnel past the active site of Ltn1, which ubiquitylates lysine residues in the emerging nascent chain[10] and with the aid of Cdc48, optimizes the conformation of Vms1 at the PTC such that it can hydrolyse the tRNA. The ubiquitylated nascent chain is engaged by Ufd1–Npl4 bound to Cdc48 that together unfold and extract the nascent chain. Dissociation of Rqc2 allows Vms1 to access the 60S subunit[29], resulting in hydrolysis of the peptidyl-tRNA. Recruitment of Cdc48–Ufd1–Npl4 to the ubiquitylated nascent chain[7] ensures its efficient extraction from the 60S subunit. It is unclear whether the action of Vms1 on 80S ribosomes is coupled in some manner to ribosome splitting.

Extended Data Table 1 | *Saccharomyces cerevisiae* cell lines and plasmids used

RJD Number	Alias	Source	Genotype
6689	S288C	Open Biosystems	*his3Δ1, leu2Δ0, met15Δ0, ura3Δ0, a*
5970	S288C pProtein A Nonstop	This Study	*his3Δ1, leu2Δ0, met15Δ0, ura3Δ0, [pGAL-ProteinA-nonstop-PGK3'UTR-URA3]. a*
6646	ubx3Δ pProtein A Nonstop	This Study	*his3Δ1, leu2Δ0, met15Δ0, ura3Δ0, ubx3::KANMX, [pGAL-ProteinA-Nonstop-PGK3'UTR-URA3], a*
6647	ubx4Δ pProtein A Nonstop	This Study	*his3Δ1, leu2Δ0, met15Δ0, ura3Δ0, ubx4::KANMX, [pGAL-ProteinA-Nonstop-PGK3'UTR-URA3], a*
6648	ubx5Δ pProtein A Nonstop	This Study	*his3Δ1, leu2Δ0, met15Δ0, ura3Δ0, ubx5::KANMX, [pGAL-ProteinA-Nonstop-PGK3'UTR-URA3], a*
6654	ufd2Δ pProtein A Nonstop	This Study	*his3Δ1, leu2Δ0, met15Δ0, ura3Δ0, ufd2::KANMX, [pGAL-ProteinA-Nonstop-PGK3'UTR-URA3], a*
6651	ufd3Δ pProtein A Nonstop	This Study	*his3Δ1, leu2Δ0, met15Δ0, ura3Δ0, ufd3::KANMX, [pGAL-ProteinA-Nonstop-PGK3'UTR-URA3], a*
6640	vms1Δ pProtein A Nonstop	This Study	*his3Δ1, leu2Δ0, met15Δ0, ura3Δ0, vms1::KANMX, [pGAL-ProteinA-nonstop-PGK3'UTR-URA3], a*
5510	ltn1Δ pProtein A Nonstop	This Study	*his3Δ1, leu2Δ0, met15Δ0, ura3Δ0, ltn::KANMX, [pGAL-ProteinA-nonstop-PGK3'UTR-URA3], a*
6670	ufd1-1 pProtein A Nonstop	This Study	*his4-519, ura3-52, ade1-100, leu2-3, 112, ufd1-1, TRP+, pGAL-ProteinA-nonstop-PGK3'UTR-URA3]*
6722	ltn1Δ vms1Δ pProtein A Nonstop	This Study	*his3Δ1, leu2Δ0, met15Δ0, ura3Δ0, vms1::KANMX, ltn1::KANMX, [pGAL-ProteinA-nonstop-PGK3'UTR-URA3], a*
6771	vms1Δltn1Δ pVMS1-HA	This Study	*his3Δ1, leu2Δ0, met15Δ0, ura3Δ0, vms1::KANMX, ltn1::KANMX, [YCplac111-VMS1-3HA], a*
5445	vms1 Δ	Open Biosystems	*his3Δ1, leu2Δ0, met15Δ0, ura3Δ0, vms1::KANMX, a*
5400	ltn1Δ	Open Biosystems	*his3Δ1, leu2Δ0, met15Δ0, ura3Δ0, ltn1::KANMX, a*
6698	vms1Δ pVMS1 HA	This Study	*his3Δ1, leu2Δ0, met15Δ0, ura3Δ0, vms1::KANMX, [YCplac111-VMS1-3HA], a*
6732	vms1Δ pvms1 Q295L HA	This Study	*his3Δ1, leu2Δ0, met15Δ0, ura3Δ0, vms1::KANMX, [YCPlac111-vms1-Q295LΔ-3HA], a*
6701	vms1Δ pProtein A Nonstop VMS1-HA	This Study	*his3Δ1, leu2Δ0, met15Δ0, ura3Δ0,vms1::KANMX, [pGAL-ProteinA-nonstop-PGK3'UTR-URA3], [YCplac111-VMS1-3HA], a*
6705	vms1Δ pProtein A Nonstop vms1-Q295L-3HA	This Study	*his3Δ1, leu2Δ0, met15Δ0, ura3Δ0,vms1::KANMX, [pGAL-ProteinA-nonstop-PGK3'UTR-URA3], [YCplac111-vms1-Q295L-3HA], a*
6733	vms1Δ pvms1 Q295P	This Study	*his3Δ1, leu2Δ0, met15Δ0, ura3Δ0, vms1::KANMX, [YCPlac111-vms1-Q295PΔ-3HA], a*
6731	vms1Δ pVMS1 ASΔ 3HA	This Study	*his3Δ1, leu2Δ0, met15Δ0, ura3Δ0, vms1::KANMX, [YCPlac111-vms1-ActiveSiteΔ-3HA], a*
6740	vms1Δ pvms1Δ622-625 (vimΔ) 3HA	This Study	*his3Δ1, leu2Δ0, met15Δ0, ura3Δ0,vms1::KANMX, [YCplac111-vms1-Δ622-625-3HA], a*
6739	vms1Δ pvms1ΔIntR 3HA	This Study	*his3Δ1, leu2Δ0, met15Δ0, ura3Δ0,vms1::KANMX, [YCplac111-vms1-ΔIntR-3HA], a*
6738	vms1Δ pvms1RR313/314AA	This Study	*his3Δ1, leu2Δ0, met15Δ0, ura3Δ0,vms1::KANMX, [YCplac111-vms1-RR313/314-3HA], a*
6707	vms1Δ pProtein A Nonstop pvms1 ASΔ 3HA	This Study	*his3Δ1, leu2Δ0, met15Δ0, ura3Δ0,vms1::KANMX, [pGAL-ProteinA-nonstop-PGK3'UTR-URA3], [YCplac111-vms1-ASΔ-3HA], a*
6706	vms1Δ pProtein A Nonstop pvms1-Q295P-3HA	This Study	*his3Δ1, leu2Δ0, met15Δ0, ura3Δ0,vms1::KANMX, [pGAL-ProteinA-nonstop-PGK3'UTR-URA3], [YCplac111-vms1-Q295P-3HA], a*

Plasmid	Relevant Vector	Source
RDB2764	pGAL-ProteinA-nonstop-PGK3'UTR[URA3]	Ambro VanHoof
RDB3226	YCplac111-VMS1-3HA	Alexander Buchberger
RDB3293	YCplac111-vms1 ΔAA74-102-HA	This Study
RDB3295	YCplac111-vms1 ΔInt R-HA	This Study
RDB3286	YCplac111-vms1 ASΔ-HA	This Study
RDB3287	Ycplac-111-vms1 Q295L-HA	This Study
RDB3288	YCplac111-vms1 Q295P-HA	This Study
RDB3292	YCplac111-vms1 RR313/314AA	This Study
RDB2632	pET28a-VMS1	Hai Rao
RDB3336	pET28a-vms1 Q295L	This Study
RDB3353	pCMV6-ANKZF1-mycDDK	Origene
RDB3359	pCMV6-ankzf1Q246L-mycDDK	This Study
RDB3299	T7-3xFLAG-CRP	This Study

Top, list of yeast strains used in this study; bottom, list of plasmids used in this study.

https://doi.org/10.1038/s41586-018-0077-3

Molecular mechanism of GPCR–mediated arrestin activation

Naomi R. Latorraca[1,2,3,4], Jason K. Wang[2], Brian Bauer[5], Raphael J. L. Townshend[2], Scott A. Hollingsworth[1,2,3], Julia E. Olivieri[4], H. Eric Xu[6,7], Martha E. Sommer[5]* & Ron O. Dror[1,2,3,4]*

Despite intense interest in discovering drugs that cause G-protein-coupled receptors (GPCRs) to selectively stimulate or block arrestin signalling, the structural mechanism of receptor-mediated arrestin activation remains unclear[1,2]. Here we reveal this mechanism through extensive atomic-level simulations of arrestin. We find that the receptor's transmembrane core and cytoplasmic tail—which bind distinct surfaces on arrestin—can each independently stimulate arrestin activation. We confirm this unanticipated role of the receptor core, and the allosteric coupling between these distant surfaces of arrestin, using site-directed fluorescence spectroscopy. The effect of the receptor core on arrestin conformation is mediated primarily by interactions of the intracellular loops of the receptor with the arrestin body, rather than the marked finger-loop rearrangement that is observed upon receptor binding. In the absence of a receptor, arrestin frequently adopts active conformations when its own C-terminal tail is disengaged, which may explain why certain arrestins remain active long after receptor dissociation. Our results, which suggest that diverse receptor binding modes can activate arrestin, provide a structural foundation for the design of functionally selective ('biased') GPCR-targeted ligands with desired effects on arrestin signalling.

G-protein-coupled receptors (GPCRs) represent the largest class of drug targets, and their interactions with arrestins are emerging as a focal point for drug discovery efforts. Arrestins not only block interactions of GPCRs with G proteins, but also promote GPCR internalization and mediate a variety of G-protein-independent signalling pathways. The design of 'biased' drugs that cause GPCRs to favour or avoid stimulation of arrestins relative to stimulation of G proteins could lead to more effective and safer treatments for a wide range of diseases[1,2].

Arrestin activation has long been associated with binding of the phosphorylated cytoplasmic tail of the receptor (R_P tail), which displaces the long C-terminal tail of arrestin (C tail)[3,4]. Gurevich and Benovic originally proposed that arrestin activation requires interaction of the receptor with two separate sites on arrestin—a 'phosphorylation sensor' and an 'activation sensor'—with arrestin undergoing an activating conformational change only when both sites are engaged[4,5]. Subsequent studies, including the recently reported structure of rhodopsin bound to arrestin-1[6,7], confirmed that the receptor binds arrestin at two distinct interfaces: the R_P tail binds within a positively charged trench in the arrestin N domain, whereas the transmembrane helices and loops of the receptor (the receptor core) bind between the N domain and C domain (Fig. 1a). Multiple studies have concluded, however, that binding of the receptor core is not required for arrestin activation, and that the R_P tail is able to stimulate arrestin conformational change and signalling on its own[8–13].

To determine whether the receptor core has a role in arrestin activation—and, more generally, to determine the structural mechanisms by which receptor binding at the R_P tail and core interfaces brings about conformational change in arrestin—we performed extensive all-atom molecular dynamics simulations of arrestins, either free or bound to various parts of the receptor. Comparison of crystal structures of arrestin indicates that the largest conformational change arrestin undergoes upon activation is a twist of the C domain of about 20° relative to the N domain[7,9,14] (Fig. 1b). We used this twist angle as a primary metric of arrestin activation in simulations, but also examined other, more localized conformational changes associated with arrestin activation.

In simulation, active arrestin-1 (also known as visual arrestin) was stable when bound to co-crystallized rhodopsin, maintaining an average interdomain twist angle of 20° (Fig. 1b, Extended Data Fig. 1). Full-length inactive arrestin-1, with its C tail bound to its N domain, was also stable, although its twist angle increased slightly from the crystallographic conformation upon loss of crystal packing contacts.

By contrast, simulations of arrestin-1 with its C tail removed—mimicking the naturally occurring p44 splice variant, which lacks the C tail—fluctuated between the active and inactive conformations with no receptor present (Fig. 1c, Extended Data Figs. 1, 2). Simulations initiated from the inactive crystal structure, but with the C tail deleted, reached active conformations with interdomain twist angles that matched those in the active crystal structure. Likewise, simulations initiated from the active structure—that is, after removing rhodopsin from the rhodopsin–arrestin complex, which lacks a resolved arrestin C tail—reached inactive conformations. This mobility between inactive and active conformations is consistent with the fact that p44 has been crystallized in both active and inactive conformations[15,16]. Simulations initiated in either conformation adopted interdomain twist angles closer to that of the crystallographic inactive conformation approximately 70% of the time (Fig. 1d), suggesting that displacement of the arrestin C tail increases arrestin's conformational mobility but leaves it predominantly in an inactive conformation.

To investigate the contributions of receptor core and R_P tail interactions to arrestin activation, we initiated simulations from the crystal structure of the rhodopsin–arrestin-1 complex, removing either the R_P tail (leaving only the receptor core) or the receptor core (leaving only the R_P tail). In each case, the C domain remained close to its active-state position, although not as consistently as when the entire receptor was present (average interdomain twist angles of 15.6 ± 5.1° with only the receptor core bound and 17.0 ± 5.1° with only the R_P tail bound, compared to 20.6 ± 4.4° with both bound and 8.7 ± 6.4° with neither present) (Fig. 1d).

These results suggest that the receptor core and R_P tail each individually stabilize the active conformation of arrestin, even though they bind at completely different interfaces, and binding of both together stabilizes it further. The simulated conformational ensembles of arrestin bound to the receptor core, the R_P tail or full-length rhodopsin overlap substantially (Extended Data Fig. 2). Indeed, we were able to

---

[1]Biophysics Program, Stanford University, Stanford, CA, USA. [2]Department of Computer Science, Stanford University, Stanford, CA, USA. [3]Departments of Molecular and Cellular Physiology and Structural Biology, Stanford University School of Medicine, Stanford, CA, USA. [4]Institute for Computational and Mathematical Engineering, Stanford University, Stanford, CA, USA. [5]Institut für Medizinische Physik und Biophysik (CC2), Charité-Universitätsmedizin Berlin, Berlin, Germany. [6]VARI-SIMM Center, Center for Structure and Function of Drug Targets, CAS-Key Laboratory of Receptor Research, Shanghai Institute of Materia Medica, Chinese Academy of Sciences, Shanghai, China. [7]Laboratory of Structural Sciences, Center for Structural Biology and Drug Discovery, Van Andel Research Institute, Grand Rapids, MI, USA. *e-mail: martha.sommer@charite.de; ron.dror@stanford.edu

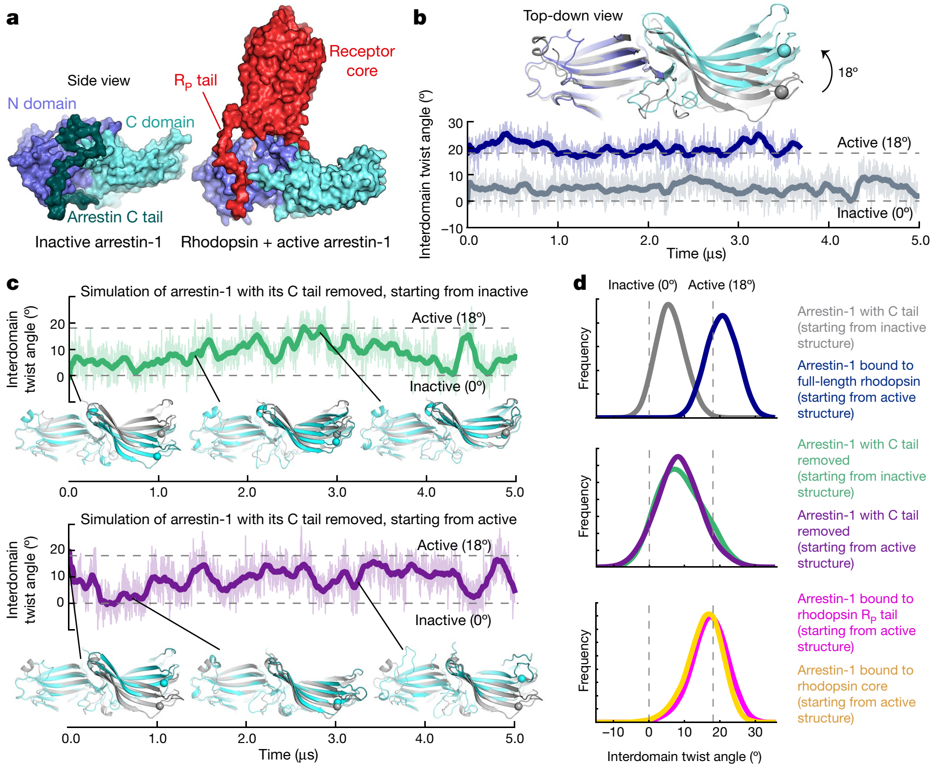


Fig. 1 | Removal of the C tail of arrestin leads arrestin to fluctuate between active and inactive states; the receptor core and R_P tail each independently stabilize the active state. a, Inactive-state (left; PDB 1CF1) and receptor-bound, active-state arrestin-1 (right; PDB 5W0P). **b**, Upon activation, the arrestin C domain twists with respect to the N domain. In simulations of arrestin-1 starting from its inactive state, the interdomain twist angle remained close to 0° (grey trace), while in simulations starting from its active state with rhodopsin bound, the twist angle remained close to 20° (blue trace). Thick traces represent a 50-ns sliding mean and thin traces represent unsmoothed values. **c**, In simulations of arrestin without its C tail and without a receptor, arrestin spontaneously transitioned between inactive and active conformations. Top, simulation of arrestin-1 with C tail removed, starting from inactive structure; bottom, arrestin-1 with C tail removed, starting from active structure. **d**, Distributions (histograms) of interdomain twist angles under different simulation conditions: grey, arrestin-1 with C tail (starting from inactive structure); blue, arrestin-1 bound to full-length rhodopsin (starting from active structure); green, arrestin-1 with C tail removed (starting from inactive structure); purple, arrestin-1 with C tail removed (starting from active structure); magenta, arrestin-1 bound to rhodopsin R_P tail (starting from active structure); yellow, arrestin-1 bound to rhodopsin core (starting from active structure). Removal of the arrestin C tail leads to an increased range of interdomain twist angles (the standard deviation increases; $P < 0.001$, two-sided t-test with six simulations per condition, see Methods). Binding of either part of the receptor in simulation substantially increases the fraction of time that arrestin spends in active conformations (histograms, yellow and magenta; $P = 0.002$ for core binding; $P = 0.003$ for R_P tail binding). Binding of the entire receptor has an even stronger effect (histogram, blue; $P = 0.01$ compared to both core-bound and to R_P-tail-bound conditions). Histograms are based on all production simulations. Traces are shown for representative trajectories; all are shown in Extended Data Fig. 1.

trigger transitions to the active state by applying force to arrestin loops that contact either the receptor core or the R_P tail, as described below. We also initiated unbiased simulations from an inactive arrestin conformation, with and without the receptor core present (see Methods). The core had an immediate activating effect on the arrestin interdomain twist angle, suggesting that the core energetically disfavours inactive-like conformations as opposed to simply slowing transitions from active to inactive conformations (Extended Data Fig. 3).

How does the receptor core trigger arrestin activation? Three elements of arrestin interact with the core in the rhodopsin–arrestin-1 structure (Fig. 2a): the arrestin finger loop inserts into the receptor's helical bundle, the C loop in the C domain of arrestin and nearby residues in the N domain interact with the second intracellular loop (IL2) of the receptor, and R318 in the back loop of the C domain forms an ionic interaction with E239 in the third intracellular loop (IL3) of the receptor.

Of these elements, the finger loop undergoes by far the largest conformational change upon receptor binding, and this conformational change has therefore been suggested to trigger arrestin activation[14,17]. Our simulations indicate, however, that the conformation of the finger loop depends almost entirely on whether or not the receptor core is bound, with little connection to the activation state of arrestin (Fig. 2b). Moreover, mimicking the effect of receptor binding by pulling the finger loop into its receptor-bound, helical conformation had little effect on interdomain twisting (Fig. 2c, Extended Data Fig. 4). These results suggest that although the finger loop is likely to contribute substantially to arrestin–GPCR binding, it plays a limited role in triggering arrestin activation (Extended Data Fig. 5).

By contrast, the conformations of arrestin regions contacted by IL2 and IL3—closer into the 'body' of arrestin—are coupled to the global conformational state of arrestin. In simulations, when arrestin shifted from an inactive state to an active state in the absence of a GPCR, the C loop tended to move away from the nearby N domain and the back loop tended to extend away from the arrestin body (Extended Data Fig. 6). Pulling the C-loop and N-domain residues contacted by IL2 into the conformation stabilized by receptor binding caused the interdomain twisting associated with arrestin activation (Fig. 2d, Extended Data Fig. 4). The receptor core thus appears to trigger arrestin activation through interactions of receptor intracellular loops with the arrestin body. These results also suggest a possible explanation for the observation that in certain GPCRs, phosphorylated or acidic residues on IL3 promote arrestin binding and activation[18]: these negatively charged residues might interact with positively charged residues on the back loop, which are found in all arrestin isoforms.

How does the R_P tail favour arrestin activation, and how does the arrestin C tail prevent it? Both bind in the same groove in the N domain

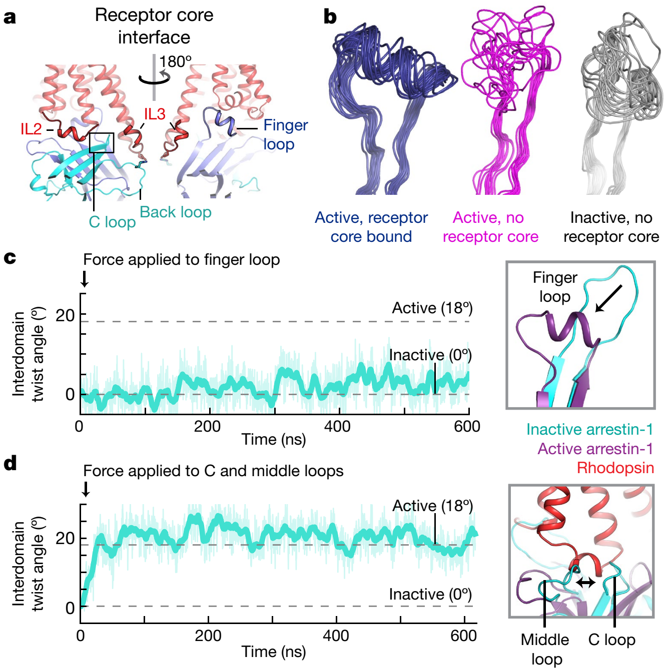


Fig. 2 | The receptor core favours arrestin activation through interactions mediated by the receptor's intracellular loops. a, The core interface. **b**, The conformation of the finger loop is largely determined by the presence or absence of a receptor, with little coupling to the arrestin activation state. In simulations of arrestin-1 bound to the receptor core, the finger loop retains a helical conformation (blue). Upon removal of the core but in the presence of the R_P tail, the finger loop collapses towards a set of disordered states (magenta), even as the global arrestin conformation remains active. This ensemble more closely resembles that observed in simulations starting from the inactive state (grey). **c**, Pulling the finger loop to its receptor-bound (helical) conformation using targeted molecular dynamics (TMD) does not induce twist angles characteristic of active states. **d**, In simulations in which the C loop is pushed away from the middle loop of the N domain, mimicking IL2 binding, arrestin adopts active conformations (six independent simulations for each condition; $P = 0.002$, two-sided *t*-test, when compared to unbiased simulations; and $P = 0.0004$ when compared to simulations in which the finger loop is pulled to its active state; see Methods). See Extended Data Fig. 4 for all TMD traces.

of arrestin, adjacent to the gate loop of the C domain. Our simulations suggest that, despite their opposing effects, both the R_P tail and the C tail influence arrestin activation by favouring particular conformations of the gate loop. The C tail is coupled to the gate loop through a network of highly conserved ionic interactions known as the polar core[19], including residue D296 at the base of the loop (Fig. 3a, b, left). Our simulations indicate that the position of D296 is tightly coupled to interdomain twisting; when D296 moves in the direction of the C domain, the adjoining β-strand (residues 292–296) moves with it, pushing against the C domain and causing it to twist (Fig. 3b, c, Extended Data Fig. 7). Removal of the C tail disrupts the polar core and allows the gate loop to flip to an alternative conformation, thus allowing D296 and the entire C domain to fluctuate between their inactive and active positions (Fig. 3c). In simulations in which the R_P tail is bound, an interaction between phosphoserine pS338 and K300 stabilizes a particular gate loop conformation that shifts D296 away from the R_P-tail-binding groove (Fig. 3a). Indeed, pulling the gate loop into the conformation favoured by R_P tail binding caused the interdomain twisting motion associated with arrestin activation (Fig. 3d, Extended Data Fig. 4).

Our computational results imply that the two GPCR-binding interfaces of arrestin are allosterically coupled to one another. In particular, our simulations indicate that both motion of the gate loop (at the R_P-tail-binding interface) and motion of the C loop (at the receptor-core-binding interface) are coupled to interdomain twisting. Motion of the C loop leads to interdomain twisting and thus to motion in the gate loop, and vice versa.

To validate this mechanism, we used fluorescently labelled arrestin mutants that report on the position of the C loop and the gate loop of arrestin-1 (Fig. 4a). We examined the individual and combined effects of receptor core binding and R_P tail binding on arrestin conformation using light-activated, non-phosphorylated rhodopsin (Rho*); inactive, non-phosphorylated opsin (Ops); and their phosphorylated counterparts (Rho*P and OpsP). Receptor activation is required for tight core binding, and phosphorylation is required for tight tail binding[6,9].

To probe C-loop position, we used an arrestin mutant labelled with the fluorophore NBD at site 251 on the C loop. Upon arrestin activation, site 251 moves away from the nearby quenching residue Y67 in the N domain, which leads to increased fluorescence from a fluorophore placed at site 251[20] (Fig. 4b). We observed that both receptor core binding (Rho*) and R_P tail binding (OpsP) induced increases in fluorescence (1.5-fold and 1.9-fold, respectively) (Fig. 4b, Extended Data Fig. 8). Engagement of both the receptor core and R_P tail (Rho*P) induced an even larger increase in fluorescence (2.4-fold). The change in fluorescence induced by the R_P tail of OpsP is particularly notable, considering that the R_P tail binds far from the receptor core interface and the C loop.

To probe gate loop position, we used the arrestin mutants I299NBD/L173W and I299NBD/L173F. Upon arrestin activation, the NBD fluorophore at site 299 in the gate loop moves close to site 173 in a nearby β-sheet (Fig. 4c). By comparing the changes in fluorescence with a quenching tryptophan or a non-quenching phenylalanine at site 173, we can calculate a quenching ratio (F_{Phe}/F_{Trp}), which increases as the gate loop moves to the active conformation[16]. The quenching ratio increased substantially not only upon binding to OpsP ($F_{Phe}/F_{Trp} = 2.4$) and Rho*P ($F_{Phe}/F_{Trp} = 2.7$), which include an R_P tail, but also upon binding to the receptor core only (Rho*, $F_{Phe}/F_{Trp} = 2.2$). This observation demonstrates that binding of the receptor core induces movement of the distant gate loop to its active conformation. Our experimental results, taken together, confirm our computational finding that receptor core and R_P tail binding can each independently cause the global conformational changes in arrestin required for activation.

Our study focused on arrestin-1, the arrestin isoform for which the most structural data are available. The very high level of structural homology between different arrestins suggests a similar mechanism for β-arrestins[5]. We performed simulations of arrestin-2 (also known as β-arrestin-1) under several conditions, and the results are consistent with a similar activation mechanism (Extended Data Fig. 9), despite certain functional differences between arrestins[21]. A wide variety of previously published data also supports the hypothesis that the receptor core and R_P tail each independently promote activation of β-arrestins (Extended Data Table 1).

More generally, our results agree with, and shed new light on, a wealth of data spanning at least three decades (Extended Data Table 1). Several cellular-level studies reported that even in the absence of an R_P tail or other phosphorylated cytoplasmic site, stimulated GPCRs mediate processes associated with arrestin activation; we suggest that this reflects arrestin activation via the receptor core interface[22,23]. Multiple spectroscopy-based studies identified dynamic sites at both the core and R_P tail interfaces that undergo conformational changes upon receptor binding[20,24]; our simulations now reveal which of these changes drive arrestin activation. Additionally, our observation of allosteric coupling between receptor-binding surfaces on arrestin provides a direct structural explanation for how several well-known arrestin mutations within the core-binding and R_P-tail-binding interfaces shift the conformational equilibrium of arrestin towards the active state (Extended Data Table 1, Extended Data Fig. 7).

Our results suggest that GPCRs could stimulate arrestins through at least three different binding modes: through interactions mediated by both the receptor core and the R_P tail, by the R_P tail only, or by the receptor core only. The first two modes have been observed experimentally in several recent studies[12,13]. The third mode may explain

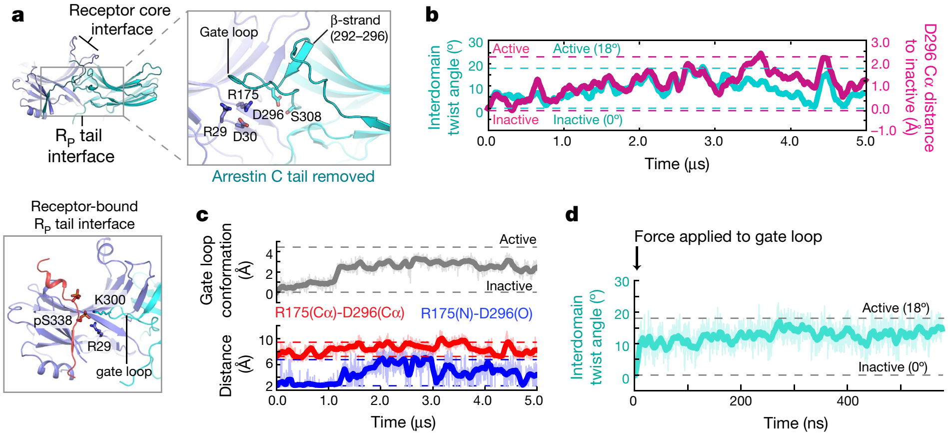


Fig. 3 | The $\mathrm{R_P}$ tail induces arrestin activation, and the arrestin C tail prevents activation, by controlling the conformation of the gate loop. **a**, Removal of the arrestin C tail from inactive arrestin frees the gate loop to adopt an alternative conformation, allowing the R175–D296 interaction to break and allowing D296 to frequent its active position (simulation snapshot, upper right). In active structures, an $\mathrm{R_P}$ tail phosphoserine engages the gate loop, stabilizing the position of D296 (bottom left). **b**, The relative position of D296 correlates with interdomain twist angle ($R^2 = 0.586$ across six simulations), as seen in a simulation starting from the inactive state with the arrestin C tail removed. **c**, Conformational changes in the gate loop towards more active-like conformations (grey trace; see Methods for description of the metric) correlate with disruption of the polar core. When the gate loop is inactive, the R175–D296 interaction (blue trace) is broken only 3.3% of the time, with the R175–D296 Cα atoms achieving active-like separation distances (red trace) only 4.0% of the time (across six independent simulations initiated from the inactive structure with the C tail removed; see Methods). By contrast, when the gate loop populates intermediate and active conformations, the R175–D296 interaction is broken 44.6% of the time, with the R175–D296 Cα atoms achieving active-like separation distances 74.1% of the time. Dashed lines indicate crystallographic positions for each metric. See Extended Data Fig. 7 for all traces. **d**, Pulling the gate loop towards the conformation observed in active-state crystal structures using TMD favours active interdomain twist angles (six independent simulations initiated from the inactive structure with the C tail removed; $P = 4 \times 10^{-6}$, two-sided *t*-test when compared to unbiased simulations; see Methods).

how some receptors lacking C-terminal phosphorylation undergo arrestin-mediated internalization[23] and how a receptor with the $\mathrm{R_P}$ tail removed can still stimulate cellular events associated with β-arrestin signalling (described in the Article by Eichel et al.[25]). These binding modes would differ in affinity, lifetime, and propensity for arrestin activation, leading to distinct cellular effects[26,27].

Recent reports have noted a curious persistence of receptor-induced conformational changes and signalling activity of β-arrestins long after

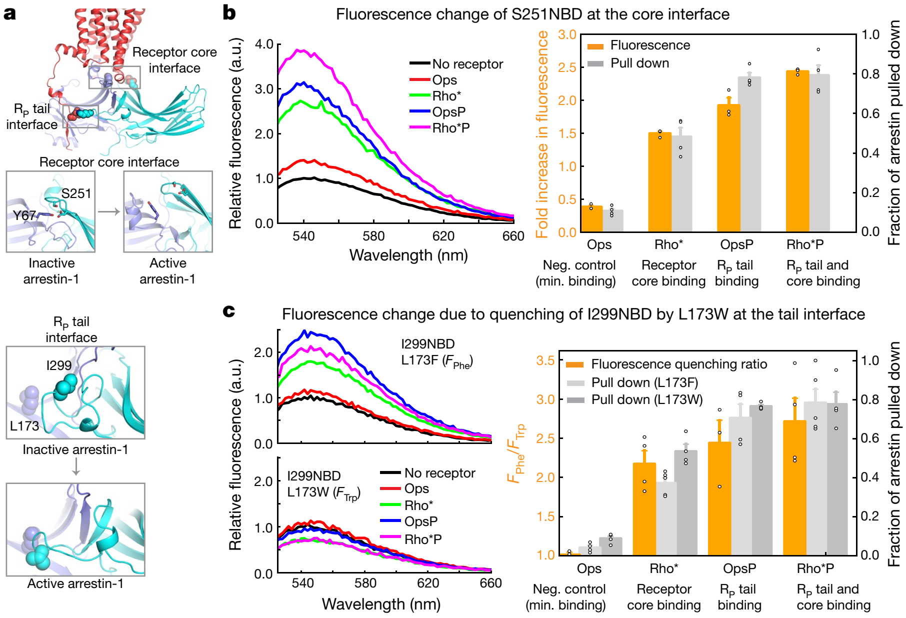


Fig. 4 | Fluorescence spectroscopy supports computational predictions. **a**, Fluorescently labelled arrestin mutants used to monitor conformational changes at the core interface (S251NBD) and $\mathrm{R_P}$ tail interface (I299NBD/L173F and I299NBD/L173W). **b**, Interdomain twisting in arrestin mutant S251NBD separates a quenching tyrosine at site 67 from the fluorophore at site 251, resulting in increased fluorescence. Increased fluorescence relative to Ops was observed in the presence of Rho* ($P = 2 \times 10^{-5}$, two-sided *t*-test), OpsP ($P = 0.005$) and Rho*P ($P = 1 \times 10^{-7}$) (orange bars show fold-increase relative to unbound condition; $n = 3$ for each condition). neg., negative; min., minimal. **c**, Disruption of the polar core is accompanied by a movement of the gate loop, which can be detected by quenching of the NBD fluorophore at site 299 by a tryptophan at site 173. As a control, site 173 is also mutated to non-quenching phenylalanine. The quenching ratio ($F_{\mathrm{Phe}}/F_{\mathrm{Trp}}$) is calculated from the spectra of I299NBD/L173F and I299NBD/L173W mutants. Substantial increases in the quenching ratio over Ops were observed in the presence of the receptor core ($P = 0.005$), $\mathrm{R_P}$ tail ($P = 0.04$) or both ($P = 0.01$), indicating that all three favoured active-like conformations of the gate loop (Rho*, Rho*P, $n = 4$; Ops, OpsP, $n = 3$). Centrifugal pull-down analysis for all mutants (grey bars) indicates that fluorescence quenching corresponded to arrestin activation and receptor binding (Rho* and Rho*P with arrestin (L173F), $n = 5$; all other mutants, $n = 4$). Fluorescence and pull-down measurements are plotted in the bar charts as mean ± s.e.m.; representative steady-state fluorescence spectra are shown.

receptor dissociation[22,28,29]. Our findings offer a potential structural explanation for this persistent arrestin-mediated signalling. As seen in our simulations, arrestin with a displaced C tail frequents active conformations even in the absence of a receptor. In β-arrestins, the displaced C tail can bind to AP2 and clathrin[30,31], slowing its re-association with the arrestin body once the receptor has dissociated.

Nevertheless, two caveats are in order. First, our simulations do not reveal the order of events during the arrestin activation process, as we have not simulated the full process of GPCR–arrestin association—particularly arrestin C-tail dissociation, which is likely to be the rate-limiting step[5,32]. Second, we have not explored the structural and functional effects of different phosphorylation patterns; this is a topic for future work.

Our results not only show that the receptor core and R_P tail can each drive arrestin activation on their own, but also suggest a key difference between how the receptor core stimulates arrestin activation and how it stimulates G protein activation. Crystal structures show that both the G protein C-terminal helix and the arrestin finger loop bind to the GPCR in a pocket between transmembrane helices 3, 5, 6 and 7, with the GPCR adopting similar conformations. Previous work has indicated that binding to this pocket triggers G protein activation[33]. Our current study suggests that the key structural changes that give rise to arrestin activation originate less in this pocket than at the interfaces of arrestin with the intracellular loops (IL2 and IL3) of the receptor. Therefore, although the receptor core adopts similar conformations in existing crystal structures of complexes with G proteins and arrestins, certain receptor core conformations might favour arrestin stimulation over G protein stimulation, or vice versa. This provides a possible avenue for the design of biased ligands.

Online content

Any Methods, including any statements of data availability and Nature Research reporting summaries, along with any additional references and Source Data files, are available in the online version of the paper at https://doi.org/10.1038/s41586-018-0077-3.

Received: 11 September 2017; Accepted: 6 March 2018;

1. Violin, J. D., Crombie, A. L., Soergel, D. G. & Lark, M. W. Biased ligands at G-protein-coupled receptors: promise and progress. *Trends Pharmacol. Sci.* **35**, 308–316 (2014).
2. DeWire, S. M. & Violin, J. D. Biased ligands for better cardiovascular drugs: dissecting G-protein-coupled receptor pharmacology. *Circ. Res.* **109**, 205–216 (2011).
3. Palczewski, K., Buczyłko, J., Imami, N. R., McDowell, J. H. & Hargrave, P. A. Role of the carboxyl-terminal region of arrestin in binding to phosphorylated rhodopsin. *J. Biol. Chem.* **266**, 15334–15339 (1991).
4. Gurevich, V. V. & Benovic, J. L. Visual arrestin interaction with rhodopsin. Sequential multisite binding ensures strict selectivity toward light-activated phosphorylated rhodopsin. *J. Biol. Chem.* **268**, 11628–11638 (1993).
5. Gurevich, V. V. & Gurevich, E. V. The structural basis of arrestin-mediated regulation of G-protein-coupled receptors. *Pharmacol. Ther.* **110**, 465–502 (2006).
6. Kang, Y. et al. Crystal structure of rhodopsin bound to arrestin by femtosecond X-ray laser. *Nature* **523**, 561–567 (2015).
7. Zhou, X. E. et al. Identification of phosphorylation codes for arrestin recruitment by G protein-coupled receptors. *Cell* **170**, 457–469 (2017).
8. Nobles, K. N., Guan, Z., Xiao, K., Oas, T. G. & Lefkowitz, R. J. The active conformation of beta-arrestin1: direct evidence for the phosphate sensor in the N-domain and conformational differences in the active states of beta-arrestins1 and -2. *J. Biol. Chem.* **282**, 21370–21381 (2007).
9. Shukla, A. K. et al. Structure of active β-arrestin-1 bound to a G-protein-coupled receptor phosphopeptide. *Nature* **497**, 137–141 (2013).
10. Kumari, P. et al. Functional competence of a partially engaged GPCR-β-arrestin complex. *Nat. Commun.* **7**, 13416 (2016).
11. Kumari, P. et al. Core engagement with β-arrestin is dispensable for agonist-induced vasopressin receptor endocytosis and ERK activation. *Mol. Biol. Cell* **28**, 1003–1010 (2017).
12. Thomsen, A. R. B. et al. GPCR–G protein–β-arrestin super-complex mediates sustained G protein signaling. *Cell* **166**, 907–919 (2016).
13. Cahill, T. J. III et al. Distinct conformations of GPCR–β-arrestin complexes mediate desensitization, signaling, and endocytosis. *Proc. Natl Acad. Sci. USA* **114**, 2562–2567 (2017).
14. Chen, Q. et al. Structural basis of arrestin-3 activation and signaling. *Nat. Commun.* **8**, 1427 (2017).
15. Granzin, J. et al. Crystal structure of p44, a constitutively active splice variant of visual arrestin. *J. Mol. Biol.* **416**, 611–618 (2012).
16. Kim, Y. J. et al. Crystal structure of pre-activated arrestin p44. *Nature* **497**, 142–146 (2013).
17. Sommer, M. E., Farrens, D. L., McDowell, J. H., Weber, L. A. & Smith, W. C. Dynamics of arrestin-rhodopsin interactions: loop movement is involved in arrestin activation and receptor binding. *J. Biol. Chem.* **282**, 25560–25568 (2007).
18. Tobin, A. B. G-protein-coupled receptor phosphorylation: where, when and by whom. *Br. J. Pharmacol.* **153** (Suppl. 1), S167–S176 (2008).
19. Hirsch, J. A., Schubert, C., Gurevich, V. V. & Sigler, P. B. The 2.8 Å crystal structure of visual arrestin: a model for arrestin's regulation. *Cell* **97**, 257–269 (1999).
20. Sommer, M. E., Smith, W. C. & Farrens, D. L. Dynamics of arrestin–rhodopsin interactions: arrestin and retinal release are directly linked events. *J. Biol. Chem.* **280**, 6861–6871 (2005).
21. Sensoy, O., Moreira, I. S. & Morra, G. Understanding the differential selectivity of arrestins toward the phosphorylation state of the receptor. *ACS Chem. Neurosci.* **7**, 1212–1224 (2016).
22. Jung, S. R., Kushmerick, C., Seo, J. B., Koh, D. S. & Hille, B. Muscarinic receptor regulates extracellular signal regulated kinase by two modes of arrestin binding. *Proc. Natl Acad. Sci. USA* **114**, E5579–E5588 (2017).
23. Jala, V. R., Shao, W. H. & Haribabu, B. Phosphorylation-independent β-arrestin translocation and internalization of leukotriene B_4 receptors. *J. Biol. Chem.* **280**, 4880–4887 (2005).
24. Yang, F. et al. Phospho-selective mechanisms of arrestin conformations and functions revealed by unnatural amino acid incorporation and (19)F-NMR. *Nat. Commun.* **6**, 8202 (2015).
25. Eichel, K. et al. Catalytic activation of β-arrestin by GPCRs. *Nature* https://doi.org/10.1038/s41586-018-0079-1 (2018).
26. Lee, M. H. et al. The conformational signature of β-arrestin2 predicts its trafficking and signalling functions. *Nature* **531**, 665–668 (2016).
27. Tohgo, A. et al. The stability of the G protein-coupled receptor-beta-arrestin interaction determines the mechanism and functional consequence of ERK activation. *J. Biol. Chem.* **278**, 6258–6267 (2003).
28. Eichel, K., Jullié, D. & von Zastrow, M. β-Arrestin drives MAP kinase signalling from clathrin-coated structures after GPCR dissociation. *Nat. Cell Biol.* **18**, 303–310 (2016).
29. Nuber, S. et al. β-Arrestin biosensors reveal a rapid, receptor-dependent activation/deactivation cycle. *Nature* **531**, 661–664 (2016).
30. Goodman, O. B. Jr et al. β-arrestin acts as a clathrin adaptor in endocytosis of the β_2-adrenergic receptor. *Nature* **383**, 447–450 (1996).
31. Laporte, S. A., Miller, W. E., Kim, K. M. & Caron, M. G. β-Arrestin/AP-2 interaction in G protein-coupled receptor internalization: identification of a β-arrestin binging site in β_2-adaptin. *J. Biol. Chem.* **277**, 9247–9254 (2002).
32. Pulvermüller, A. et al. Functional differences in the interaction of arrestin and its splice variant, p44, with rhodopsin. *Biochemistry* **36**, 9253–9260 (1997).
33. Dror, R. O. et al. Structural basis for nucleotide exchange in heterotrimeric G proteins. *Science* **348**, 1361–1365 (2015).

Acknowledgements We thank P. Eastman, M. Sultan, R. Betz and C. Brinton for assistance with simulation setup and analysis, and K. Eichel, M. Masureel, A. Venkatakrishnan, B. Kobilka and M. von Zastrow for valuable discussions. This work was supported by National Institutes of Health (NIH) grant R01GM127359 to R.O.D., the Deutsche Forschungsgemeinschaft (SO1037/1-2 to M.E.S.), the Berlin Institute of Health (Delbrück Fellowship BIH_PRO_314 to M.E.S.), an NIH postdoctoral fellowship to S.A.H. (T15-LM007033-33), a Stanford Terman Faculty Fellowship to R.O.D., and NSF Graduate Research Fellowships to N.R.L. and R.J.L.T. A few preliminary simulations were performed on an Anton machine at the Pittsburgh Supercomputing Center, donated by D.E. Shaw Research and supported by NIH grant R01GM116961.

Reviewer information *Nature* thanks C. Hoffmann and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions N.R.L., M.E.S. and R.O.D. designed the research. N.R.L and J.K.W. performed and analysed simulations, with assistance from R.J.L.T., J.E.O., S.A.H. and R.O.D. B.B. prepared arrestin mutants. M.E.S. performed and analysed fluorescence spectroscopy experiments. H.E.X. provided structural information and insights. N.R.L, M.E.S. and R.O.D. wrote the paper.

Competing interests The authors declare no competing interests.

Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0077-3.
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0077-3.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to M.E.S. or R.O.D.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

No statistical methods were used to predetermine sample size. The experiments were not randomized and investigators were not blinded to allocation during experiments and assessment.

System setup for molecular dynamics simulations. We simulated arrestin-1 under six conditions, initiated from the following structures: (1) the inactive-state crystal structure of bovine arrestin-1 bound to the crystallographic fragment of its C tail (PDB entry 1CF1, chain A, residues 10–362 and 374–393)[19]; (2) the same inactive-state crystal structure with the C tail removed (residues 374–393); (3) the active-state crystal structure of arrestin-1 bound to full-length rhodopsin with residues Thr336 and Ser338 phosphorylated (PDB entry 5W0P)[7]; (4) the same active-state crystal structure of arrestin-1 bound to the receptor core of rhodopsin only (residues 1–324); (5) the same active-state crystal structure of arrestin-1 bound to the rhodopsin R_P tail only (residues 325–344, with Thr336 and Ser338 phosphorylated; these simulations used an earlier refinement of the now-published 5W0P crystal structure, available upon request); and (6) an active-state structure of arrestin that was determined in complex with rhodopsin (PDB entry 4ZWJ)[6], but with the entire rhodopsin molecule removed for simulation. For simplicity, we refer to the co-crystallized receptor, a constitutively active mutant of opsin, as 'rhodopsin' throughout our manuscript. Residue numbers refer to bovine rhodopsin and arrestin-1.

We also performed simulations of arrestin-2 under four conditions, initiated from the following structures: (1) the inactive-state arrestin-2 crystal structure bound to the crystallographic fragment of its C tail (PDB entry 1G4M, chain A, residues 5–361 and 372–386)[34]; (2) the same inactive-state crystal structure with the C tail removed; (3) the active-state arrestin-2 crystal structure bound to the V2 vasopressin receptor (V2R) R_P tail (PDB entry 4JQI)[9]; and (4) the active-state arrestin-2 crystal structure with the V2R R_P tail removed. For simulations starting from the active state of arrestin-2, we removed the co-crystallized Fab30 antibody. Simulations containing rhodopsin's receptor core included a lipid bilayer; in these simulations, palmitoyl modifications, which help to anchor helix 8 of rhodopsin in the membrane, were added to residues 322 and 323. In some simulations, we restrained the residues in the finger loop to examine whether maintaining the finger loop in its receptor-bound helical conformation favoured the active state. In other simulations, we used TMD to pull the conformations of various local regions in arrestin, including the finger loop, the gate loop and the C loop, from their inactive crystallographic conformation to their receptor-bound crystallographic conformation. We performed multiple simulations for each condition (Supplementary Table 1). For each simulation, initial atom velocities were assigned randomly and independently.

Simulation coordinates were prepared by removing non-protein molecules from all initial crystal structures. Prime (Schrödinger) was used to model missing side chains and loops, and neutral acetyl and methylamide groups were added to cap protein termini. In arrestin simulations, we retained titratable residues in their dominant protonation state at pH 7. In rhodopsin simulations, Asp83, Glu122 and Glu134 were protonated, in accordance with evidence that these residues become protonated upon activation[35]. Constitutively activating mutations present in the rhodopsin–arrestin-1 crystal structure, including N2C/N282C and E113Q/M257Y, were retained in these simulations to ensure that rhodopsin retained an active conformation while bound to arrestin. Histidines were represented with hydrogen on the epsilon nitrogen except in cases where addition of hydrogen to the delta nitrogen helped to optimize the local hydrogen bond network.

For simulations performed in the presence of a bilayer, the prepared protein structures were aligned on the transmembrane helices to the 'orientation of proteins in membranes' (OPM)[36] structure for PDB 4ZWJ. The aligned structures were then inserted into either a pre-equilibrated palmitoyl-oleoyl-phosphatidylcholine (POPC) bilayer or a box of water (for simulations of arrestin alone) using in-house simulation preparation software[37]. Sodium and chloride ions were added to neutralize each system at a concentration of 150 mM. Water-box dimensions were chosen to maintain at least an 18 Å buffer between protein images in all dimensions. Bilayer dimensions were chosen to maintain at least a 35 Å buffer between protein images in the x–y plane and a 20 Å buffer between protein images in the z direction. System dimensions are listed in Supplementary Table 1 for each simulation condition.

Simulation protocols. For all simulations, we used the CHARMM36 force field for proteins, lipids and ions and the TIP3P model for waters[38–42]. We generated parameters for the custom residue palmitoylcysteine using the Paramchem webserver[43–45].

We performed the majority of simulations using the Compute Unified Device Architecture (CUDA) version of Particle-Mesh Ewald Molecular Dynamics (PMEMD) in AMBER on one or two graphical processing units (GPUs)[46]. Simulations were performed using the AMBER15 and AMBER16 software[47]. Systems were heated from 0 K to 100 K in the NVT ensemble over 12.5 ps and then from 100 K to 310 K in the NPT ensemble over 125 ps, using 10.0 kcal mol^{-1} Å^{-2} harmonic restraints applied to lipid and protein heavy atoms. Systems were then equilibrated at 310 K in the NPT ensemble at 1 bar, with harmonic restraints on all protein heavy atoms tapered off by 1.0 kcal mol^{-1} Å^{-2} starting at 5.0 kcal mol^{-1} Å^{-2} in a stepwise fashion every 2 ns for 10 ns and then by 0.1 kcal mol^{-1} Å^{-2} in a stepwise fashion every 2 ns for 20 ns. Production simulations were performed in the NPT ensemble at 310 K and 1 bar, using a Langevin thermostat for temperature coupling and a Monte Carlo barostat for pressure coupling. These simulations used a 4-fs time step with hydrogen mass repartitioning[48]. Bond lengths to hydrogen atoms were constrained using SHAKE. Non-bonded interactions were cut off at 9.0 Å, and long-range electrostatic interactions were computed using particle mesh Ewald (PME) with an Ewald coefficient of approximately 0.31 Å and an interpolation order of 4. The FFT grid size was chosen such that the width of a grid cell was approximately 1 Å. Trajectory snapshots were saved every 200 ps.

Prior to performing the extensive set of AMBER simulations described above, we performed three simulations of arrestin-1 on the Anton machine at the Pittsburgh Supercomputing Center (simulations 10, 11 and 39 in Supplementary Table 1)[49]. Systems first underwent the equilibration protocol described above in AMBER15 with a 2.5-fs time step and no hydrogen mass repartitioning. We then transferred equilibrated systems to Anton and performed production runs of 6–8 μs in length with a RESPA integrator, which employed a time step of 2.3333 fs and calculated long-range electrostatics every three time steps, or once every 7 fs. These simulations employed u-series electrostatics (C. Predescu et al., unpublished). Trajectory snapshots were saved every 240 ps. These initial simulations motivated our further exploration of arrestin-1 conformational dynamics because they established that arrestin exhibits extreme conformational flexibility in the absence of its C tail, starting from either active or inactive conformations.

TMD and restrained molecular dynamics simulations were conducted on GPUs using the OpenMM molecular simulation platform with the Plumed 2.01 plug-in to introduce the additional energy term[50–52]. In all TMD conditions, the force constant k was increased linearly from 0 to 20 kcal mol^{-1} Å^{-2} during the first 10 ns of production and then held constant at 20 kcal mol^{-1} Å^{-2} for the remainder of the simulation, as done in a previous study[53]. Simulations ran for 500–600 ns. For simulations involving the gate loop (residues 295–306) and finger loop (residues 67–80), we applied the pulling force to backbone Cα atoms and aligned the simulated structure to the reference structure on the N domain of arrestin. For the C loop, we aligned simulated structures on only the residues to which we also applied the force—namely, residues within contact of the rhodopsin IL2 on the middle loop (residues 132–143) and C loop (residues 245–257) in the recent rhodopsin–arrestin crystal structure (PDB entry 5W0P). For traces that display interdomain twisting in TMD simulations, we display the 5 ns of equilibration before application of the pulling force.

In order to obtain a conformation of arrestin in an inactive state with the receptor core present, we performed twenty independent TMD simulations to pull the arrestin to the inactive state in the presence of the receptor core (after removing the R_P tail). As above, forces were applied to backbone Cα atoms within the β-sheets of the C domain that were not in contact with the receptor core in the rhodopsin–arrestin crystal structure (residues 202–210, 220–228, 234–242, 260–264, 272–280, 328–336 and 344–352) and increased linearly over 100 ns of simulation to 10.0 kcal mol^{-1} Å^{-2}. We chose a simulation frame that exhibited the minimum root mean square deviation (r.m.s.d.) from the inactive state crystal structure (PDB 1CF1, chain A), calculated on the atoms used for pulling, from 20 independent TMD simulations (r.m.s.d. = 0.856 Å). This initial inactive conformation possessed at least three characteristics of inactive states: in addition to the low interdomain twist angle, the gate loop exists in a configuration it can spontaneously adopt in simulations of arrestin with just the C tail removed, and the polar core interaction between R175 and D296 is reformed.

In order to examine the effect of the receptor core on the inactive state of arrestin-1, we then initiated 20 production simulations with randomized velocities from this snapshot and from the same snapshot but with the receptor core removed in AMBER. For the simulations with receptor core removed, we prepared the system in a water box and performed a shortened equilibration protocol, as described above (with restraints tapered every 1 ns, rather than every 2 ns), before performing the production simulations. Not all production simulations were of the same length owing to limited computing resources.

Analysis protocols for molecular dynamics simulations. The AmberTools15 CPPTRAJ package was used to reimage and centre trajectories[54]. Simulations were visualized and analysed using Visual Molecular Dynamics (VMD)[55]. Time traces from simulation were smoothed using a moving average with a window size of 50 ns unless otherwise indicated and visualized with the PyPlot package from Matplotlib. For all analysis in the manuscript that required structural alignment, including calculation of interdomain twist angle and projection metric (described below), we aligned arrestin structures on the N domain of inactive arrestin-1 (residues 11–179, PDB entry 1CF1).

Throughout the manuscript, we report an interdomain twist angle. This corresponds to how much the C domain rotates away from its position in the inactive crystal structure relative to the N domain. In order to capture rotation towards more active-like or inactive-like conformations, we measure rotation about a particular axis: the rotation axis that defines the displacement of the C domain (relative to the N domain) between the inactive and active arrestin-1 crystal structures. Thus the inactive crystal structure has an interdomain twist angle of 0°; positive interdomain twist angles represent motion in the direction seen in the active crystal structure; and negative interdomain twist angles represent motion in the opposite direction. For calculations in Extended Data Fig. 9, we replace the arrestin-1 inactive state structure with the inactive arrestin-2 crystal structure (PDB entry 1G4M). We use the rhodopsin-bound arrestin-1 structure as the active state structure for these calculations, as this is the only available structure of an arrestin bound to a receptor core.

To estimate distributions of interdomain twist angles for each simulation condition ('histograms'), we used a Gaussian kernel density estimator from Python's scikit-learn package with a covariance factor of 0.25°.

To quantify the position and conformation of the gate loop, finger loop or back loop in a given simulation frame (Fig. 3, Extended Data Figs. 1, 4, 6, 7), we used a metric that specifies how 'active-like' or 'inactive-like' the loop's conformation is. In particular, we represent the positions of all gate loop (or finger or back loop) C-alpha atoms as a single vector containing their Cartesian coordinates; this might be thought of as a point in a $3n$-dimensional space, where n is the number of gate loop (or finger or back loop) residues. The crystallographic inactive and active conformations of the gate loop (or finger or back loop) can also each be represented as a point in this $3n$-dimensional space. We project the vector representing the gate loop (or finger or back loop) in a given simulation frame onto the line connecting the crystallographic active and inactive conformations. We then report the position of the projected point on that line, using the convention that the inactive conformation is at 0 and positive values indicate change in the direction of the active conformation. The result is that more active-like conformations will be assigned larger values than more inactive-like conformations. In particular, the crystallographic inactive conformation will be assigned a value of 0 Å, and the crystallographic active conformation will be assigned a value equal to its r.m.s.d. from the inactive conformation. We performed a similar analysis in which we aligned on the N domain and computed the projection metric on the entire C domain (Extended Data Fig. 1).

For principal component analysis (PCA), we applied the PCA class in scikit-learn to the Cartesian coordinates of Cα atoms for residues 11–360 across all six simulation conditions for arrestin-1 after aligning every simulation on the N domain (residues 11–179) of the inactive-state crystal structure, using simulation trajectories downsampled every 20 ns (for computational efficiency), after removing the first 500 ns in order to reduce the effects of initial transients (that is, achieve better equilibration).

To test the significance of differences in trajectories between simulations performed under different conditions, we used a two-sided Welch's t-test, treating each independent simulation as a separate data point. We performed six production simulations under each condition ($n = 6$), excluding the first 500 ns of each from the analysis in order to reduce the effects of initial transients (that is, achieve better equilibration). We compared standard deviations of interdomain twist angle to test claims about differences in ranges of these angles. We compared means of interdomain twist angle to test claims about differences in extent of activation.

In Fig. 1d (top and middle panels), we tested the hypothesis that removing the arrestin C tail increased the range of interdomain twist angles visited in simulation. Removing the arrestin C tail increased the range of interdomain twist angles visited in simulation for simulations starting from the inactive-state structure (green versus grey histograms; $P = 0.0007$). We also found that simulations starting from arrestin's active-state structure in the absence of both the receptor and the C tail visited an increased range of interdomain twist angles compared to simulations starting from the inactive-state structure in the presence of arrestin's C tail (purple versus grey histograms; $P = 0.0002$).

In Fig. 1d (top and bottom panels), we also found that binding of both the receptor core and R_P tail (dark blue histogram) together led to an increase in the average interdomain twist angle sampled in simulation compared to simulations in which only the receptor core was bound (yellow histogram; $P = 0.005$) or in which only the R_P tail was bound (magenta histogram; $P = 0.01$).

To determine the fraction of the time the polar core exhibited active-like conformations when the gate loop moved towards the active state, we treated the gate loop as inactive if its conformation had deviated less than 1 Å from the inactive state and active-like if its conformation had deviated at least 2 Å from the inactive state, towards the active state. We treated the polar core ionic interaction as active if the side chain polar atoms were separated by at least 4 Å and the polar core Cα distance as active if the Cα atoms came within 0.5 Å of the active-state separation distance (approximately the standard deviation of that metric in simulation of the active-state crystal structure). These calculations refer to the condition in which arrestin was simulated from its inactive state with its C tail removed.

Preparation of opsin and rhodopsin. Bovine retinas were extracted from eyes obtained from a local abattoir under dim red light and frozen in 45% buffered sucrose solution. Rod outer segments (ROS) were isolated from these retinas under dim red light using discontinuous sucrose-gradient centrifugation as previously described[56]. Briefly, 100 to 200 thawed retinas were shaken to break off outer segments, and the suspension was centrifuged (2,500g, 5 min). The supernatant was filtered through cheesecloth, diluted 1:1 slowly with ROS buffer (70 mM potassium phosphate, 1 mM $MgCl_2$, 0.1 mM EDTA, 1 mM DTT, 0.5 mM PMSF, pH 7), and then centrifuged (6,000g, 10 min). The pellets were gently homogenized in 25.5% sucrose (1.105 g ml^{-1}) and then layered onto four gradients composed of 27.125% (1.115 g ml^{-1}) sucrose underlaid with 32.25% (1.135 g ml^{-1}) sucrose. The gradients were centrifuged in a swinging bucket rotor at 83,000g for 30 min. ROS were collected at the 1.115 g ml^{-1}–1.135 g ml^{-1} interface and then snap frozen in liquid N_2.

For rhodopsin phosphorylation, ROS were thawed and diluted about 1:1 in ROS buffer under dim red light. ATP (8 mM), $MgCl_2$ (1 mM) and 11-*cis*-retinal (10–20 μM) were added. Note that 11-*cis*-retinal was purified by HPLC using commercially available all-*trans*-retinal as the starting reagent[57]. The ROS suspension was divided between Falcon tubes and placed on a rocking platform under a standard desk lamp at room temperature. After 2 h, 50 mM NH_2OH was added to convert all rhodopsin photoproducts to opsin. ROS were then washed several times by pelleting the membranes by centrifugation followed by resuspension in a generous volume of potassium phosphate buffer (100 mM, pH 7). Membranes were then washed once more in 50 mM HEPES pH 7. The collected pellets were resuspended in a small volume of HEPES buffer, aliquoted, and snap frozen in liquid N_2. Note that non-phosphorylated membrane preparations were treated in exactly the same way, except that ATP was omitted before light-exposure of the ROS.

Rhodopsin was regenerated by the addition of a threefold molar excess of 11-*cis*-retinal to opsin membranes (> 1-h incubation in the dark, room temperature). Regeneration was terminated by the addition of NH_2OH, which was then washed away using the wash protocol described above. Rhodopsin concentration was determined by the loss of 500-nm absorbance (extinction coefficient 0.0408 $\mu M^{-1} cm^{-1}$) after bleaching (> 495 nm) the membranes diluted 1:20 in 100 mM NH_2OH. High levels of phosphorylation in phosphorylated rhodopsin preparations were confirmed using the 'extra meta II' assay[56,58] and isoelectric focusing[59].

Preparation of fluorescently labelled arrestin mutants. Mutations were introduced into a 'base construct' consisting of bovine arrestin-1, with an additional glycine inserted after the first methionine, and the following mutations: C63A, C128S, C143A and W194F. This recombinant arrestin gene was cloned into the pET15b vector for bacterial expression[60]. For this study, the S251C, I299C/L173F and I299C/L173W mutants were used, which were previously created for other studies[16,61]. Arrestin mutants were expressed and purified by two-step ion-exchange chromatography as described[60]. Briefly, plasmid DNA was used to transform *Escherichia coli* BL21 (DE3) competent cells (New England Biolabs) and plated onto LB medium containing ampicillin. A single colony was used to inoculate 5 ml LB containing ampicillin (100 μg ml^{-1}), which was incubated for 8 h with shaking (at 28 °C). This starter culture was used to inoculate 150 ml LB plus ampicillin, which was incubated overnight with shaking (28 °C) and then split between four flasks each containing 2 l LB plus ampicillin. The cells were induced with 30 μM IPTG upon reaching an absorbance ($A_{600\,nm}$) of 0.6 and then allowed to grow for more than 16 h (28 °C) before harvesting by centrifugation. The cells were resuspended in cold lysis buffer (10 mM Tris-HCl, 2 mM EDTA, 100 mM NaCl, 5 mM DTT, pH 7.5) and lysed using a microfluidizer (Microfluidics) in the presence of DNase. The lysate was centrifuged (27,000g, 30 min) and $(NH_4)_2SO_4$ was added to the supernatant (0.32 g ml^{-1}). The precipitant was collected by centrifugation and then resuspended in 10 mM Tris-HCl, 2 mM EDTA, 5 mM DTT, pH 7. The protein suspension was cleared by centrifugation and filtration and loaded onto three 5-ml HiTrap heparin columns (GE Healthcare), while diluting 1:3 with 10 mM Tris-HCl, 2 mM EDTA, 100 mM NaCl, 5 mM DTT, pH 7. The loaded column was then washed with the same buffer, and a gradient of NaCl (0.1–0.5 M) was applied to elute the arrestin. SDS–PAGE was used to identify arrestin-containing fractions, which were pooled and filtered, and 5 mM DTT was added. This protein was loaded onto a 5-ml HiTrap SP column coupled to a HiTrap Q column (GE Healthcare) while diluting 1:10 with 10 mM Tris-HCl, 2 mM EDTA, 5 mM DTT, pH 8.5. After loading, the SP column was removed, and the Q column was washed with buffer. Arrestin was eluted with a two-step NaCl gradient; 0–0.1 M and 0.1–0.5 M. The arrestin-containing fractions were again identified by SDS–PAGE, pooled and concentrated; buffer was exchanged against isotonic buffer (50 mM HEPES, 130 mM NaCl, pH 7), aliquoted, and snap frozen in liquid N_2.

Single-cysteine arrestin mutants were labelled with the fluorophore NBD (*N*,*N*′-dimethyl-*N*-(iodoacetyl)-*N*′-(7-nitrobenz-2-oxa-1,3-diazol-4-yl) ethylenediamine) (ThermoFisher Scientific). Arrestin was diluted to 20–50 μM in isotonic buffer and NBD stock (~20 mM solubilized in DMSO) was added in sequential additions of tenfold molar excess every hour. Total DMSO concentration in the labelling reaction was maintained $< 5\%$. After 3 to 5 h, the excess label was removed by multiple concentration–dilution steps using centrifugal microconcentrators (Amicon Ultra-0.5, 10-kDa cutoff) followed by size-exclusion chromatography (Sephadex G15 from Sigma). Concentration and labelling efficiency was determined using the extinction coefficients 0.025 $\mu M^{-1} cm^{-1}$ for NBD, 0.02636 $\mu M^{-1}\ cm^{-1}$ for I299C/L173W and 0.02076 $\mu M^{-1}\ cm^{-1}$ for S251C and I299C/L173F.

Fluorescence spectroscopy. Steady-state fluorescence was measured using a SPEX Fluorolog (1680) instrument in front-face mode. The samples were excited at 500 nm, and emission was collected at 524–660 nm (2-nm step size, 0.5-s integration per point). Excitation slits were kept < 0.2 mm to minimize light-activation of rhodopsin, and emission was collected through 4-mm-wide slits. Samples (100 μl) were placed in a black-jacketed fluorescence cuvette with a small window (2 × 5 mm) and 3-mm path length (Hellma Analytics). Samples generally contained 1 μM fluorescently labelled arrestin with or without ROS membranes containing 10 μM opsin (Ops), phosphorylated opsin (OpsP), rhodopsin (Rho) or phosphorylated rhodopsin (RhoP). Opsin experiments were performed in 50 mM HEPES pH 8.5 to favour the inactive receptor core conformation[62] and promote robust arrestin association to OpsP[56]. Rhodopsin experiments were performed in isotonic buffer (50 mM HEPES, 130 mM NaCl, pH 7). Fluorescence was measured in the dark-state and after light-activation (> 495 nm, 10 s). Sigma Plot 13.0 was used for processing fluorescence spectra (for example, subtracting background fluorescence, normalizing spectra, determining integrated fluorescence intensity). Spectra were normalized to the spectrum of the unbound condition. In Fig. 4, we show fluorescence spectra for the unbound condition (black traces) obtained under isotonic conditions. The same spectra obtained under low salt conditions were nearly identical (data not shown). For the arrestin mutant I299NBD/L173W, the quenching ratio was determined by comparison of its steady-state fluorescence intensity to that of the control mutant, I299NBD/L173F. The quenching ratio is defined as F_{Phe}/F_{Trp}, where F_{Phe} is the integrated, normalized fluorescence intensity of the unquenched control, and F_{Trp} is the integrated, normalized fluorescence intensity of the tryptophan-containing mutant[63]. Each of the fluorescence quenching ratio data points in Fig. 4c is calculated as the ratio of integrated fluorescence in a single I299NBD/L173W experiment and integrated fluorescence in a single I299NBD/L173F experiment, resulting in a larger spread among the data points relative to those reported for S251NBD fluorescence. Note that for the S251NBD mutant, in which Y67 quenches the fluorophore at site 251 in the basal state of arrestin, no unquenched control is available, since replacing site 67 with phenylalanine results in protein precipitation (as previously reported[16]). Despite this, we have established that the intrinsic fluorescence changes of the labelled S251C mutant indicate loop movements in the region of the C loop relative to the base of the finger loop[16,20,61].

Centrifugal pull-down analysis. Identical samples as prepared for fluorescence experiments were subjected to pull-down analysis, to quantify the amount of arrestin bound to ROS membranes for each experiment. Briefly, 50-μl samples of the same composition as described above were centrifuged at 20,800*g* for 10 min. Rhodopsin-containing samples were light-activated (> 495 nm, 15 s) just before centrifugation. Following centrifugation, the supernatant was removed, and the pellets were solubilized in loading buffer containing 2% SDS and subjected to SDS–PAGE. Bands were visualized with Coomassie dye and gels were scanned using an Epson photo scanner. Bands were quantified using the free online program GelQuant.NET (Version 1.8.2). Binding is expressed as a percentage of the total amount of arrestin present in each experiment (2.25 μg).

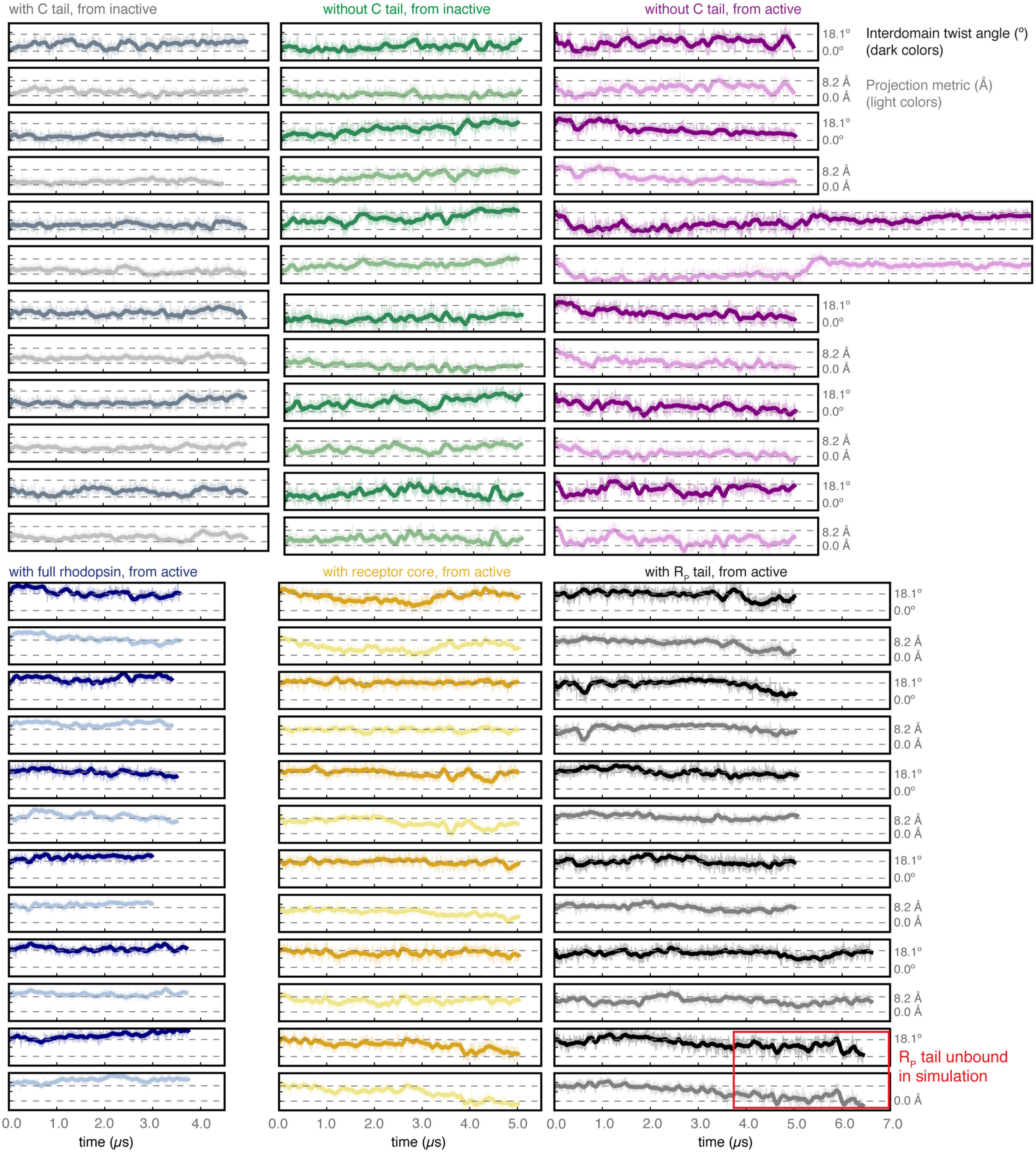
Reporting summary. Further information on experimental design is available in the Nature Research Reporting Summary linked to this paper.

Data and code availability. All simulation trajectories (Supplementary Table 1) and analysis code are available upon reasonable request to the authors, and summary statistics for all simulations are reported throughout this Letter and in its Extended Data figures.

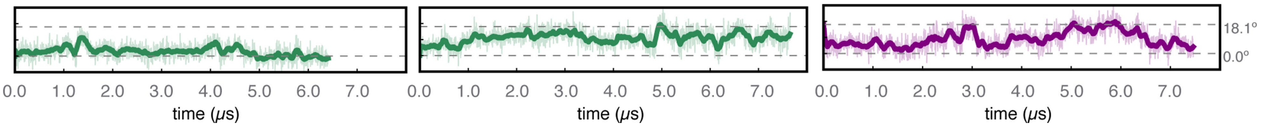
34. Han, M., Gurevich, V. V., Vishnivetskiy, S. A., Sigler, P. B. & Schubert, C. Crystal structure of β-arrestin at 1.9 Å: possible mechanism of receptor binding and membrane translocation. *Structure* **9**, 869–880 (2001).
35. Mahalingam, M., Martínez-Mayorga, K., Brown, M. F. & Vogel, R. Two protonation switches control rhodopsin activation in membranes. *Proc. Natl Acad. Sci. USA* **105**, 17795–17800 (2008).
36. Lomize, M. A., Pogozheva, I. D., Joo, H., Mosberg, H. I. & Lomize, A. L. OPM database and PPM web server: resources for position of proteins in membranes. *Nucleic Acids Res.* **40**, D370–D376 (2012).
37. Betz, R. M. Dabble. https://doi.org/10.5281/zenodo.836914 (2018).
38. MacKerell, A. D. et al. All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J. Phys. Chem. B* **102**, 3586–3616 (1998).
39. Best, R. B., Mittal, J., Feig, M. & MacKerell, A. D. Jr. Inclusion of many-body effects in the additive CHARMM protein CMAP potential results in enhanced cooperativity of α-helix and β-hairpin formation. *Biophys. J.* **103**, 1045–1051 (2012).
40. Best, R. B. et al. Optimization of the additive CHARMM all-atom protein force field targeting improved sampling of the backbone φ, ψ and side-chain χ_1 and χ_2 dihedral angles. *J. Chem. Theory Comput.* **8**, 3257–3273 (2012).
41. Huang, J. & MacKerell, A. D. Jr. CHARMM36 all-atom additive protein force field: validation based on comparison to NMR data. *J. Comput. Chem.* **34**, 2135–2145 (2013).
42. Klauda, J. B. et al. Update of the CHARMM all-atom additive force field for lipids: validation on six lipid types. *J. Phys. Chem. B* **114**, 7830–7843 (2010).
43. Vanommeslaeghe, K. et al. CHARMM general force field: a force field for drug-like molecules compatible with the CHARMM all-atom additive biological force fields. *J. Comput. Chem.* **31**, 671–690 (2010).
44. Vanommeslaeghe, K., Raman, E. P. & MacKerell, A. D. Jr. Automation of the CHARMM General Force Field (CGenFF) II: assignment of bonded parameters and partial atomic charges. *J. Chem. Inf. Model.* **52**, 3155–3168 (2012).
45. Vanommeslaeghe, K. & MacKerell, A. D. Jr. Automation of the CHARMM General Force Field (CGenFF) I: bond perception and atom typing. *J. Chem. Inf. Model.* **52**, 3144–3154 (2012).
46. Salomon-Ferrer, R., Götz, A. W., Poole, D., Le Grand, S. & Walker, R. C. Routine microsecond molecular dynamics simulations with AMBER on GPUs. 2. Explicit solvent particle mesh Ewald. *J. Chem. Theory Comput.* **9**, 3878–3888 (2013).
47. Case, D. A. et al. *AMBER 2017* (University of California, San Francisco, 2017).
48. Hopkins, C. W., Le Grand, S., Walker, R. C. & Roitberg, A. E. Long-time-step molecular dynamics through hydrogen mass repartitioning. *J. Chem. Theory Comput.* **11**, 1864–1874 (2015).
49. Shaw, D. E. et al. Millisecond-scale molecular dynamics simulations on Anton. *Proc. ACM/IEEE Conf. Supercomputing (SC09)* (2009).
50. Eastman, P. et al. OpenMM 4: A reusable, extensible, hardware independent library for high performance molecular simulation. *J. Chem. Theory Comput.* **9**, 461–469 (2013).
51. Eastman, P. & Pande, V. S. Efficient nonbonded interactions for molecular dynamics on a graphics processing unit. *J. Comput. Chem.* **31**, 1268–1272, https://doi.org/10.1002/jcc.21413 (2010).
52. Tribello, G. A., Bonomi, M., Branduardi, D., Camilloni, C. & Bussi, G. PLUMED2: New feathers for an old bird. *Comp. Phys. Comm.* **186**, 604–613 (2014).
53. Cheng, X., Wang, H., Grant, B., Sine, S. M. & McCammon, J. A. Targeted molecular dynamics study of C-loop closure and channel gating in nicotinic receptors. *PLOS Comput. Biol.* **2**, e134 (2006).
54. Roe, D. R. & Cheatham, T. E. III PTRAJ and CPPTRAJ: software for processing and analysis of molecular dynamics trajectory data. *J. Chem. Theory Comput.* **9**, 3084–3095 (2013).
55. Humphrey, W., Dalke, A. & Schulten, K. VMD: visual molecular dynamics. *J. Mol. Graph.* **14**, 27–38 (1996).
56. Sommer, M. E., Hofmann, K. P. & Heck, M. Distinct loops in arrestin differentially regulate ligand binding within the GPCR opsin. *Nat. Commun.* **3**, 995 (2012).
57. Garwin, G. G. & Saari, J. C. High-performance liquid chromatography analysis of visual cycle retinoids. *Methods Enzymol.* **316**, 313–324 (2000).
58. Schleicher, A., Kühn, H. & Hofmann, K. P. Kinetics, binding constant, and activation energy of the 48-kDa protein-rhodopsin complex by extra-metarhodopsin II. *Biochemistry* **28**, 1770–1775 (1989).
59. McDowell, J. H., Nawrocki, J. P. & Hargrave, P. A. Isolation of isoelectric species of phosphorylated rhodopsin. *Methods Enzymol.* **315**, 70–76 (2000).
60. Lally, C. C., Bauer, B., Selent, J. & Sommer, M. E. C-edge loops of arrestin function as a membrane anchor. *Nat. Commun.* **8**, 14258 (2017).
61. Sommer, M. E., Hofmann, K. P. & Heck, M. Arrestin-rhodopsin binding stoichiometry in isolated rod outer segment membranes depends on the percentage of activated receptors. *J. Biol. Chem.* **286**, 7359–7369 (2011).
62. Vogel, R. & Siebert, F. Conformations of the active and inactive states of opsin. *J. Biol. Chem.* **276**, 38487–38493 (2001).
63. Mansoor, S. E., McHaourab, H. S. & Farrens, D. L. Mapping proximity within proteins using fluorescence spectroscopy. A study of T4 lysozyme showing that tryptophan residues quench bimane fluorescence. *Biochemistry* **41**, 2475–2484 (2002).
64. Hanson, S. M. et al. A model for the solution structure of the rod arrestin tetramer. *Structure* **16**, 924–934 (2008).
65. Mukherjee, S. et al. Aspartic acid 564 in the third cytoplasmic loop of the luteinizing hormone/choriogonadotropin receptor is crucial for phosphorylation-independent interaction with arrestin2. *J. Biol. Chem.* **277**, 17916–17927 (2002).
66. Isberg, V. et al. Generic GPCR residue numbers—aligning topology maps while minding the gaps. *Trends Pharmacol. Sci.* **36**, 22–31 (2015).
67. Piana, S., Lindorff-Larsen, K. & Shaw, D. E. How robust are protein folding simulations with respect to force field parameterization? *Biophys. J.* **100**, L47–L49 (2011).
68. Vishnivetskiy, S. A. et al. How does arrestin respond to the phosphorylated state of rhodopsin? *J. Biol. Chem.* **274**, 11451–11454 (1999).
69. Bouvier, M. et al. Removal of phosphorylation sites from the β_2-adrenergic receptor delays onset of agonist-promoted desensitization. *Nature* **333**, 370–373 (1988).

70. Ohguro, H., Palczewski, K., Walsh, K. A. & Johnson, R. S. Topographic study of arrestin using differential chemical modifications and hydrogen/deuterium exchange. *Protein Sci.* **3**, 2428–2434 (1994).
71. Gurevich, V. V. & Benovic, J. L. Mechanism of phosphorylation-recognition by visual arrestin and the transition of arrestin into a high affinity binding state. *Mol. Pharmacol.* **51**, 161–169 (1997).
72. Vishnivetskiy, S. A. et al. An additional phosphate-binding element in arrestin molecule. Implications for the mechanism of arrestin activation. *J. Biol. Chem.* **275**, 41049–41057 (2000).
73. Richardson, M. D. et al. Human substance P receptor lacking the C-terminal domain remains competent to desensitize and internalize. *J. Neurochem.* **84**, 854–863 (2003).
74. Hanson, S. M. & Gurevich, V. V. The differential engagement of arrestin surface charges by the various functional forms of the receptor. *J. Biol. Chem.* **281**, 3458–3462 (2006).
75. Shukla, A. K. et al. Distinct conformational changes in β-arrestin report biased agonism at seven-transmembrane receptors. *Proc. Natl Acad. Sci. USA* **105**, 9988–9993 (2008).
76. Gimenez, L. E. et al. Role of receptor-attached phosphates in binding of visual and non-visual arrestins to G protein-coupled receptors. *J. Biol. Chem.* **287**, 9028–9040 (2012).
77. Vishnivetskiy, S. A., Baameur, F., Findley, K. R. & Gurevich, V. V. Critical role of the central 139-loop in stability and binding selectivity of arrestin-1. *J. Biol. Chem.* **288**, 11741–11750 (2013).

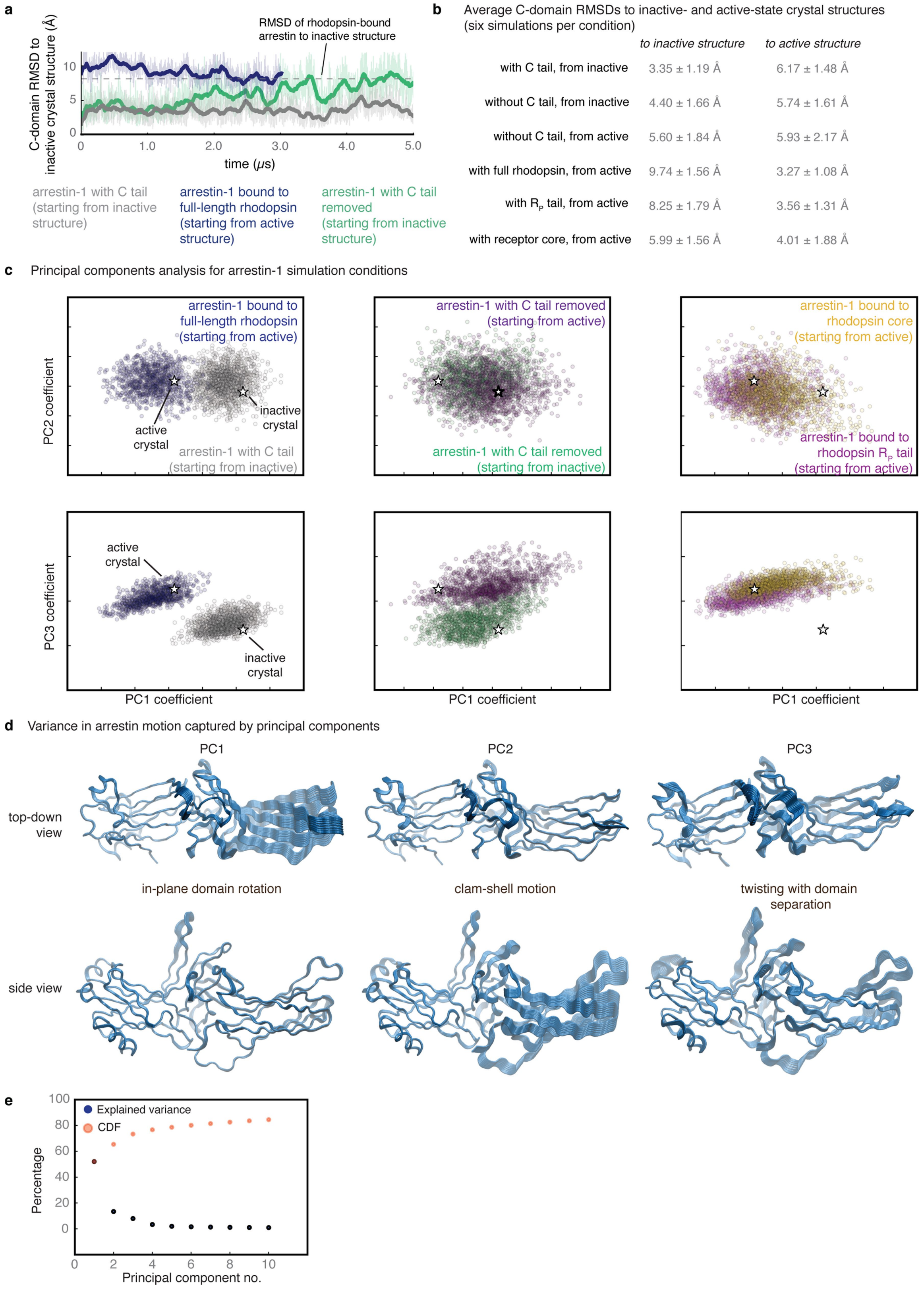
Interdomain twist angle (°) and projection metric (Å) traces across arrestin-1 simulation conditions



Interdomain twist angles in three preliminary Anton simulations



Extended Data Fig. 1 | Interdomain twist angles and global projection metric values for six arrestin-1 simulation conditions. Dashed lines indicate the interdomain twist angles in the inactive (0°) and active (18°) state crystal structures and the projection metric values in the inactive (0.0 Å) and active (8.15 Å) state structures. Thick traces indicate the moving average smoothed over a 50-ns window, and thin traces represent unsmoothed data. For each simulation, a pair of plots is shown, one immediately above the other. The top plot (dark colours) shows the interdomain twist angle. The bottom plot shows the projection metric, an alternative means of capturing global conformational change. In one simulation (red box, lower right corner), the R_P tail became unbound from arrestin, which resulted in inactive twist angles. All other simulations in that condition maintained stable binding to the R_P tail.

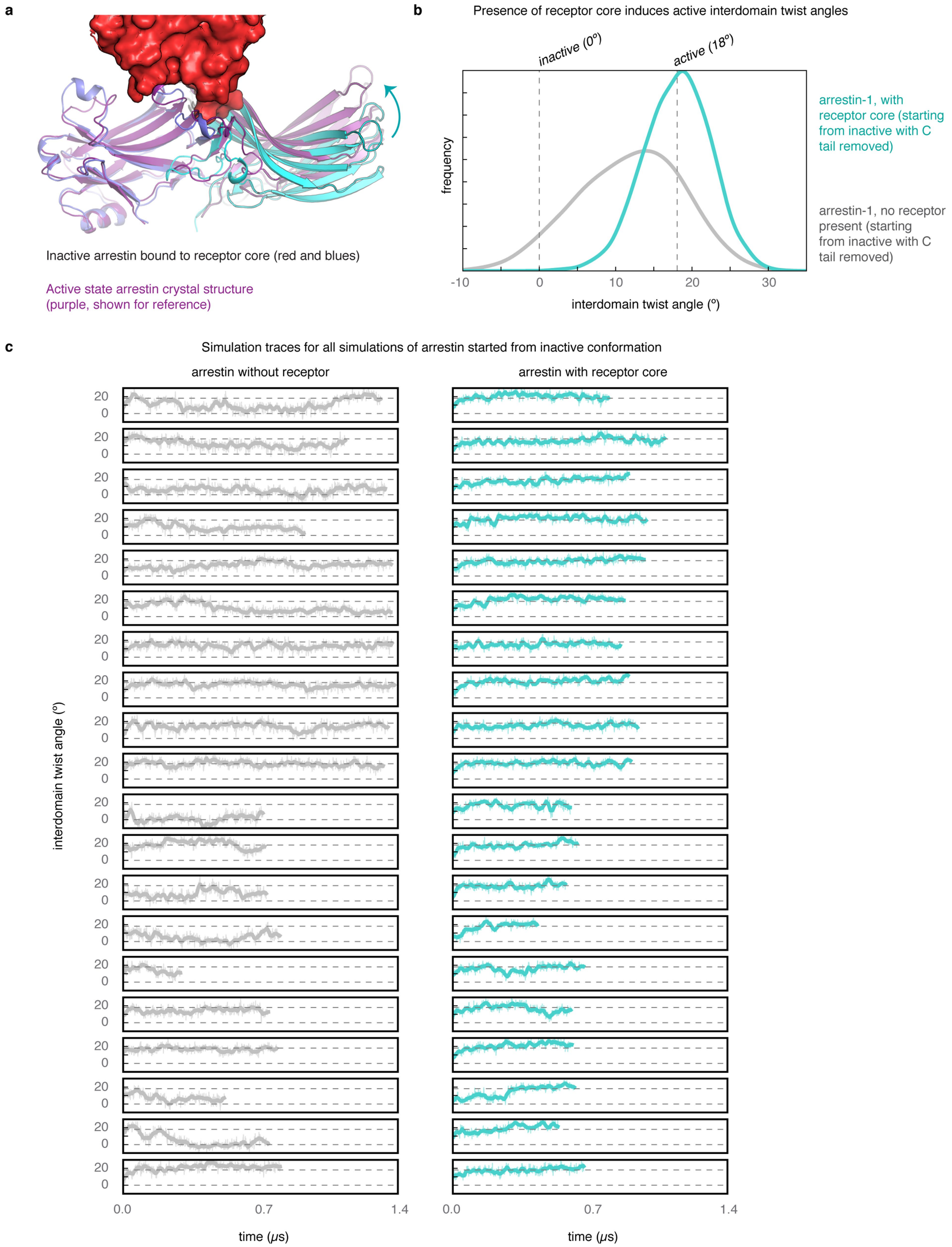


b Average C-domain RMSDs to inactive- and active-state crystal structures (six simulations per condition)

	to inactive structure	*to active structure*
with C tail, from inactive	3.35 ± 1.19 Å	6.17 ± 1.48 Å
without C tail, from inactive	4.40 ± 1.66 Å	5.74 ± 1.61 Å
without C tail, from active	5.60 ± 1.84 Å	5.93 ± 2.17 Å
with full rhodopsin, from active	9.74 ± 1.56 Å	3.27 ± 1.08 Å
with R_P tail, from active	8.25 ± 1.79 Å	3.56 ± 1.31 Å
with receptor core, from active	5.99 ± 1.56 Å	4.01 ± 1.88 Å

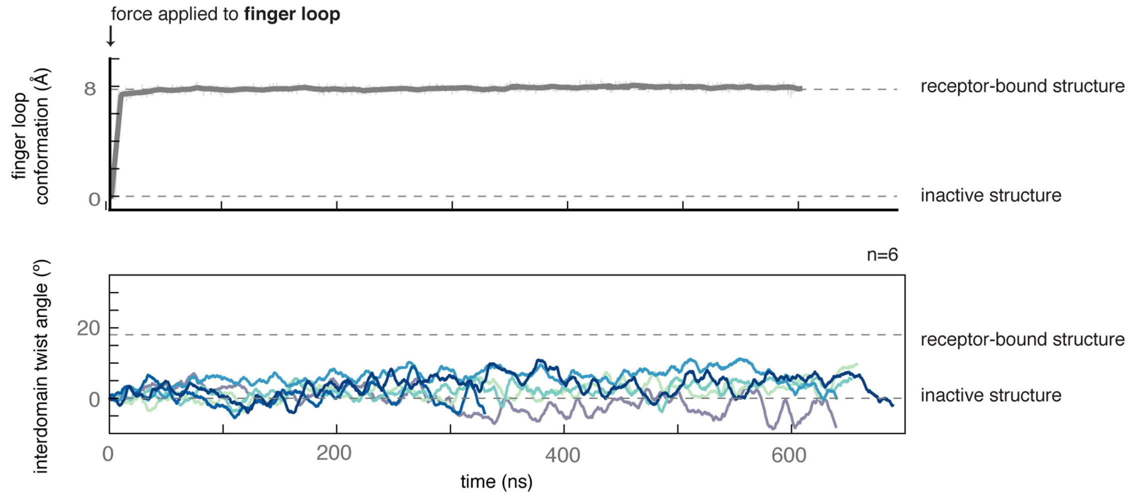
Extended Data Fig. 2 | See next page for caption.

Extended Data Fig. 2 | Global conformational behaviour of arrestin-1 in simulation. a, The r.m.s.d. from the inactive structure for representative simulations of arrestin-1 starting in its inactive state with: the C tail removed (green), arrestin-1 with the C tail present (grey), or arrestin-1 bound to full-length rhodopsin (blue). The simulation of arrestin-1 with its C tail removed transitions to active conformations and achieves r.m.s.d. values that match those of rhodopsin-bound active-state simulations. r.m.s.d. is computed on arrestin C-domain β-strands after alignment on the N domain. **b**, Mean r.m.s.d. from active and inactive structures across all six independent simulations for each condition, calculated after removing the first 500 ns of each simulation. **c**, We used PCA to compare the conformational states visited under the various arrestin-1 simulation conditions (see Methods; $n = 8100$ simulation frames as input). Each principal component corresponds to a mode of motion or variance in Cartesian coordinate space. The star on the left in each plot corresponds to the position of the active-state crystal structure, and the star on the right corresponds to the inactive-state structure. Simulations of the two crystallographic conditions separate clearly along the first principal component (PC1) and along the third principal component (PC3) but not along the second principal component (PC2). Simulations starting from the inactive state or active state with the arrestin C tail removed and no receptor present explore similar ranges of PC1 and PC2 coefficients and have some overlap in the range of PC3 coefficients. Simulations with either the receptor core or R_P tail bound closely overlap with simulations performed in the presence of the full-length receptor. The *x* axis is shifted to the right in the first plot in each row relative to the second and third plots in order to show the full range of values of PC1 coefficients. **d**, Images that show the motion of arrestin-1 along each principal component. **e**, Variance explained by each principal component. The cumulative distribution function (CDF) shows the variance explained by all principal components up to and including a given one.

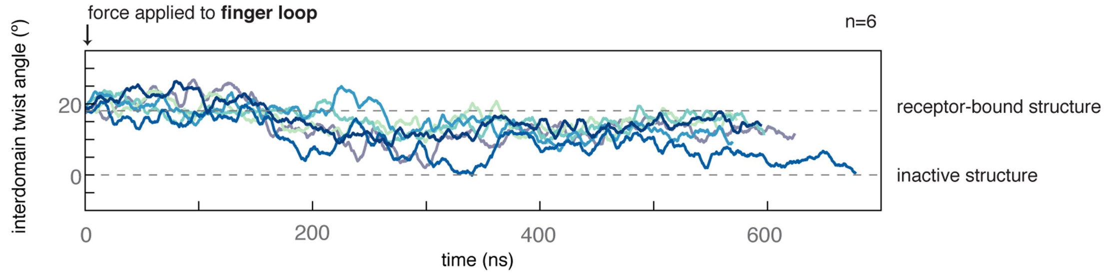


Extended Data Fig. 3 | Conformation of arrestin favoured by binding of receptor core. a, **b**, In simulations started from an inactive conformation of arrestin bound to the receptor core alone (**a**, see Methods), arrestin preferred to adopt active interdomain twist angles (cyan histogram) (**b**). By contrast, simulations of arrestin initiated from the same conformation without the receptor (grey histogram) were less likely to spontaneously adopt interdomain twist angles matching those seen in the active-state structure. **c**, Traces for all simulations, 20 per condition. The difference between the grey and cyan histograms increases with simulation time and would be likely to increase further with additional simulation time, but the differences in the mean interdomain twist angle achieved between the two conditions are already highly significant ($P = 3 \times 10^{-5}$).

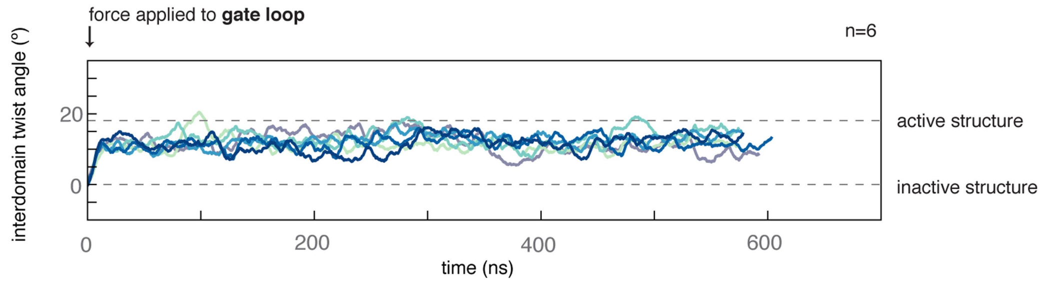
a Finger loop pulled to active state in simulations of arrestin-1 starting from inactive state



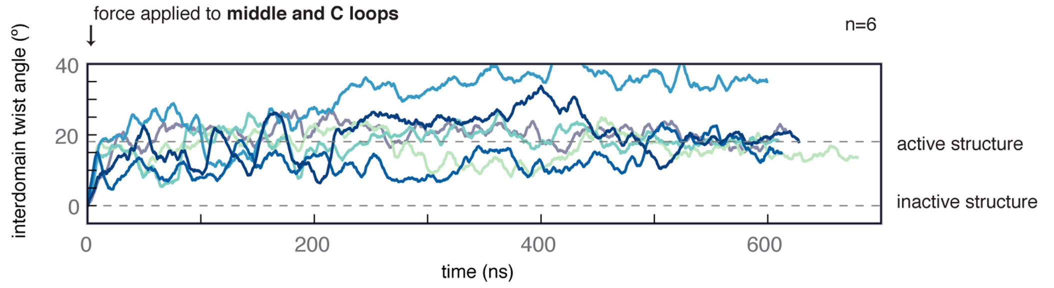
b Finger loop restrained to active state in simulations of arrestin-1 starting from active state



c Gate loop pulled to active state in simulations of arrestin-1 starting from inactive state

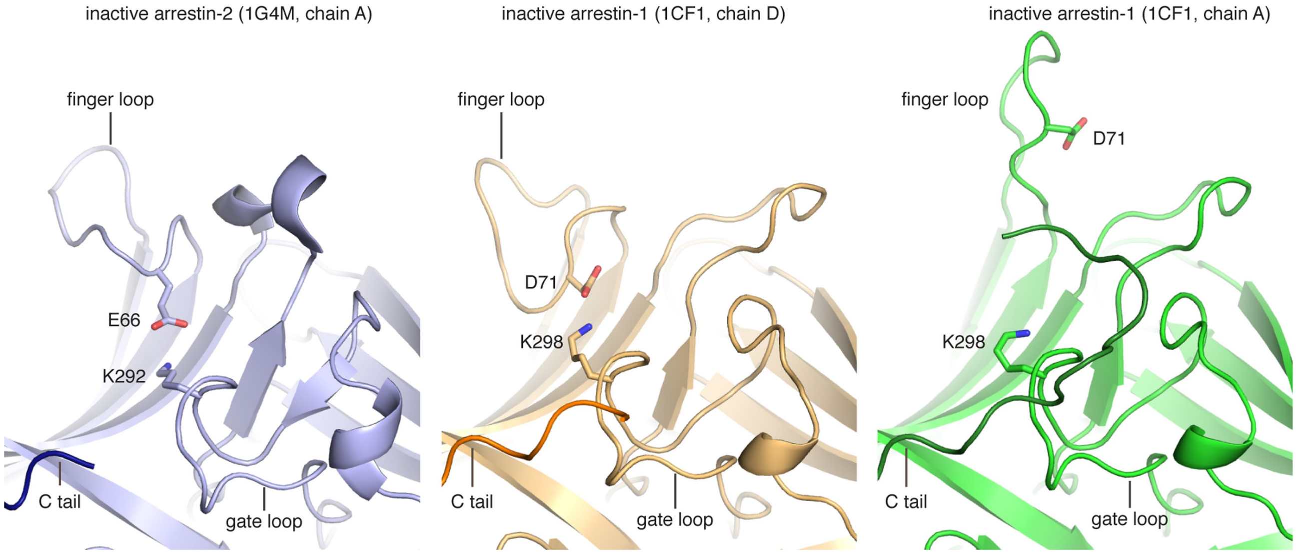


d C loop pulled to active state in simulations of arrestin-1 starting from inactive state

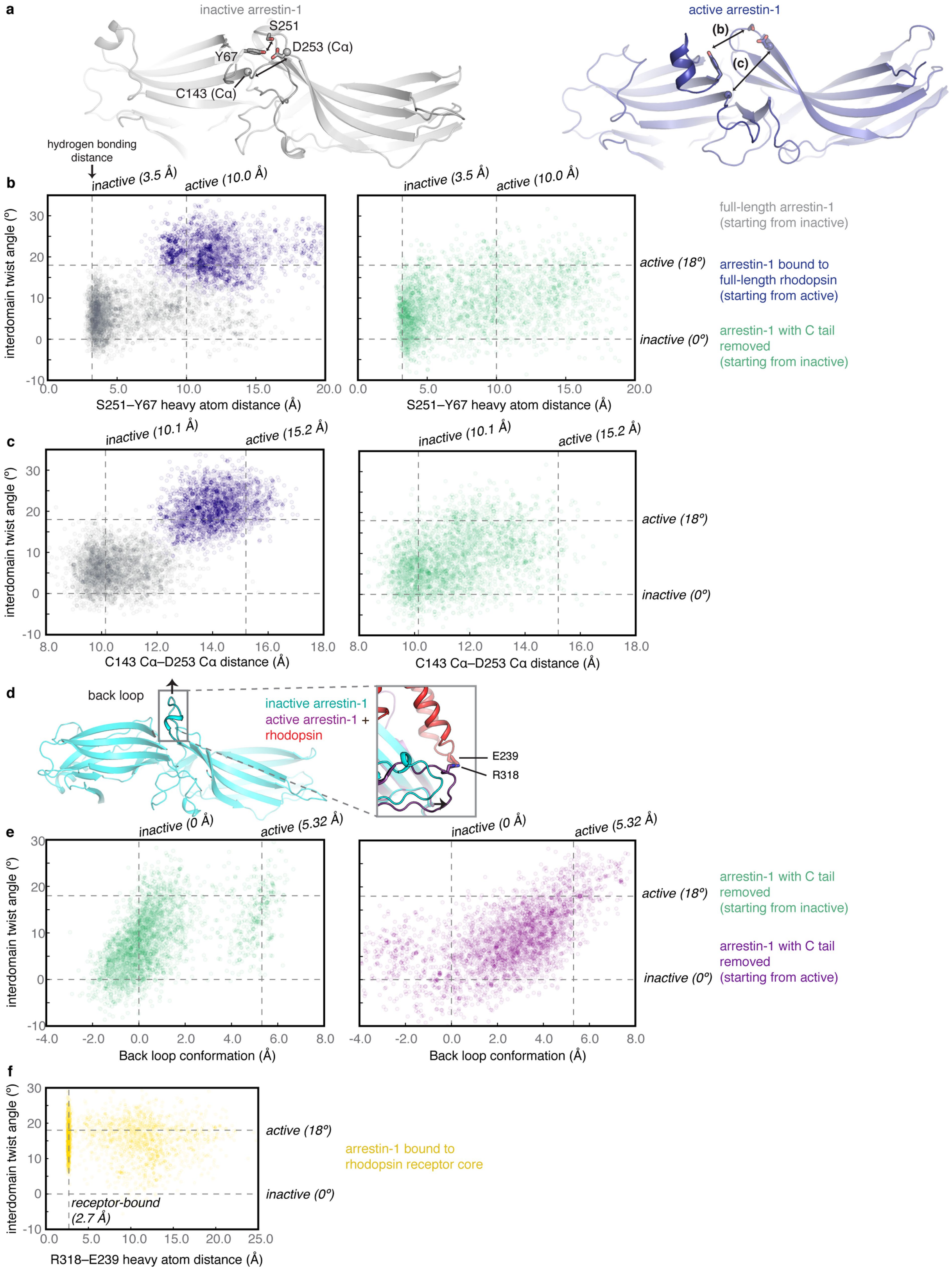


Extended Data Fig. 4 | See next page for caption.

Extended Data Fig. 4 | Simulation traces from targeted and restrained molecular dynamics simulations. **a**, We started simulations of arrestin-1 from the inactive state and pulled the finger loop towards its active, helical conformation (grey trace). The finger loop quickly reached its active state (top), but this failed to induce active interdomain twist angles (bottom) on timescales of hundreds of nanoseconds. **b**, Similarly, in simulations of arrestin-1 started from the active state (without a receptor present), restraining the finger loop conformation to its active conformation did not prevent arrestin-1 from visiting inactive interdomain twist angles. **c**, **d**, By contrast, pulling the gate loop to its active state to mimic R_P-tail binding (**c**), or pulling the IL2-binding crevice apart to mimic receptor core binding (**d**), consistently induced active interdomain twist angles in arrestin ($P = 4 \times 10^{-6}$, $P = 0.002$, respectively, compared to unbiased simulations). All traces are smoothed using an averaging window of 20 ns. In all cases we attempted to apply external forces to mimic binding of various structural elements of the receptor. To mimic the effect of binding at the receptor core interface, we aligned and pulled on the same residues within the C loop and middle loop contacted by the receptor. The broad range of interdomain twist angles may reflect the fact that the restraints do not perfectly mimic the effect of the receptor core. Nonetheless, other simulations, including unbiased simulations starting from the inactive state, suggest that separation of the interdomain crevice or the presence of the receptor core favour active interdomain twist angles (Extended Data Figs. 3, 6), providing independent support for the proposed effect of core binding on arrestin activation. Six independent simulations were performed for each condition.

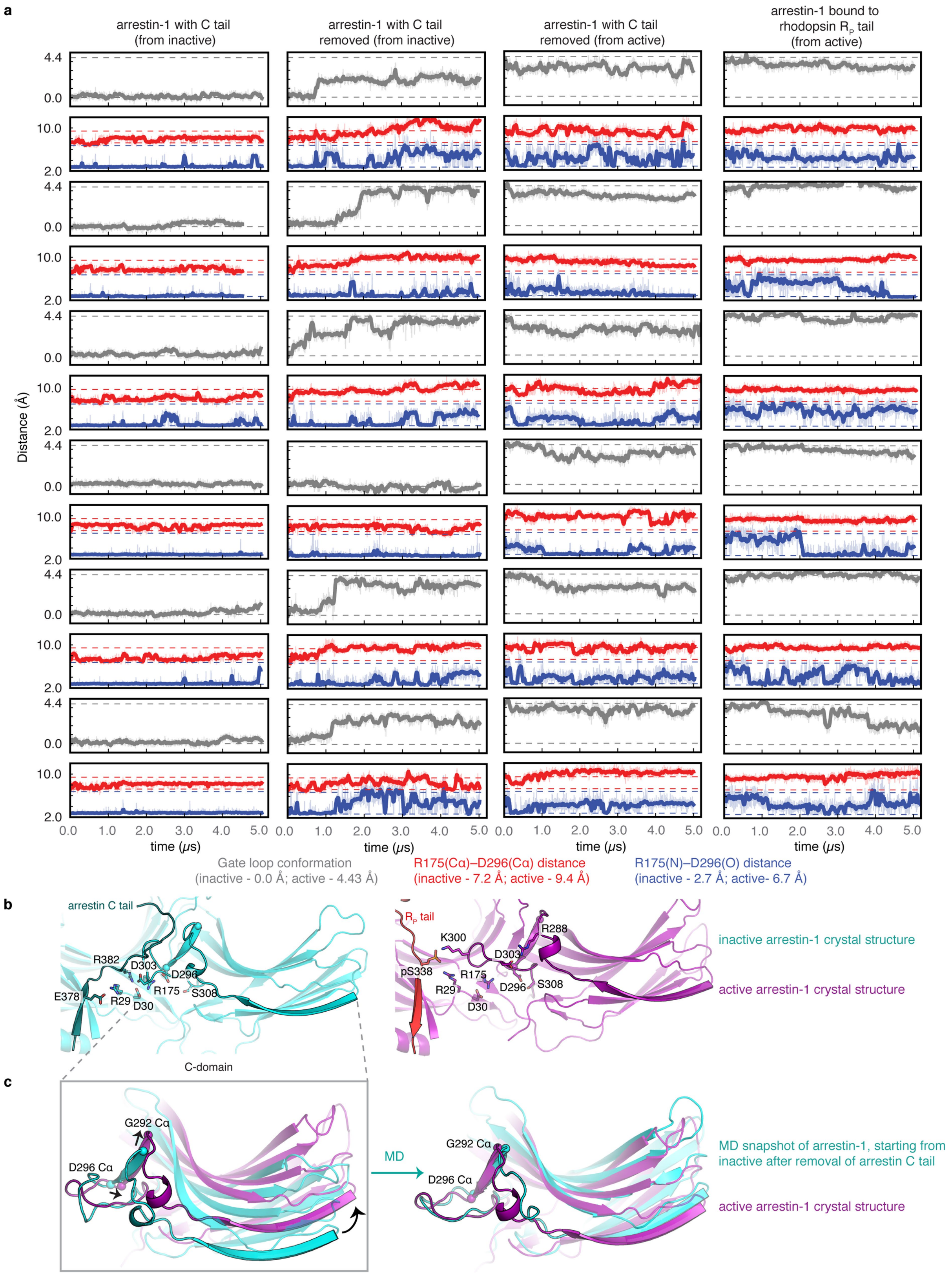


Extended Data Fig. 5 | Gate loop motion may be restrained in the inactive state by an ionic interaction with the finger loop. In certain inactive-state crystal structures of arrestin-1 (for example, PDB entry 1CF1, chain D) and arrestin-2 (for example, PDB entry 1G4M, chain A), a lysine in the gate loop (K298 in arrestin-1, K292 or K294 in arrestin-2) forms an ionic interaction with a carboxylic acid in the finger loop (D71 in arrestin-1, E66 in arrestin-2). Simulations initiated from these structures with the C tail removed exhibited less frequent transitions of the gate loop to fully active conformations than simulations initiated from crystal structures in which this ionic interaction between the gate loop and the figure loop was not formed (for example, PDB entry 1CF1, chain A). Thus a particular finger loop conformation might mildly increase the stability of the inactive-state gate loop conformation. In simulations, we observed additional sets of ionic interactions between gate loop lysines and either D67 on the finger loop or D135 on the middle loop (according to arrestin-2 numbering), which also appeared to prevent motion of the gate loop towards the active state. Certain finger loop conformations might also favour the inactive state through interactions with the C tail of arrestin[64].



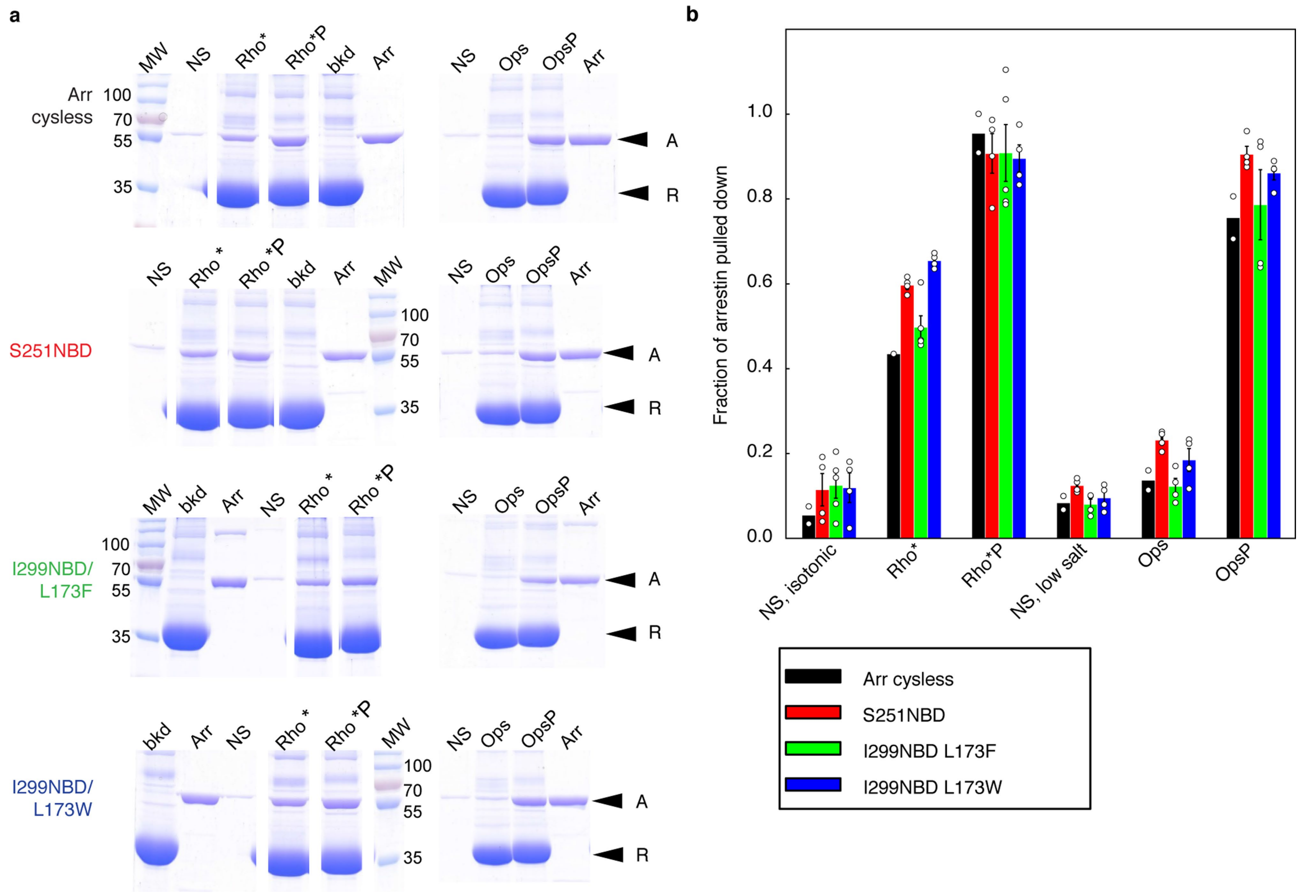
Extended Data Fig. 6 | See next page for caption.

Extended Data Fig. 6 | Conformational changes at IL2 and IL3 interfaces correlate with interdomain twist angle. a, The C loop contains residues S251 and D253, which interact with Y67 in the N domain in the inactive state of arrestin (cyan, left). In the rhodopsin-bound crystal structure, this network of residues separates when IL2 binds in the central crevice between the N and C domains (purple, right). **b**, **c**, We measured separation between the Y67–S251 side-chain hydroxyl oxygens (**b**) and between the C143–D253 Cα atoms (**c**). Conformational changes at the IL2 interface correlate with interdomain twist angles. This is particularly noticeable in simulations starting from the inactive state but with the arrestin C tail removed (green), where increased interdomain twist angles correlated with disruption of the Y67–S251 interaction ($R^2 = 0.35$) and with increased separation distance between the two domains, as measured through the C143–D253 Cα distance ($R^2 = 0.36$) (six independent simulations). Plots and correlations refer to trajectories downsampled every 10 ns, with no frames removed at the beginning of simulation. One caveat is that in simulations started from active state without the arrestin C tail, the interdomain crevice frequently collapsed at the beginning of simulation, so that even when arrestin visited more active interdomain twist angles, the crevice did not re-open. It is possible that these simulations reached a local energy minimum not typically visited in the equivalent simulations started from the inactive conformation. **d**, Conformational changes at the IL3 interface correlate with interdomain twist angles. Compared to the inactive state (blue) of arrestin, in the active state (purple), the back loop, located in the arrestin C domain (residues 311–320), extends away from the arrestin body (motion indicated by the black arrow). In this conformation, the back loop contacts the third intracellular loop in rhodopsin via an ionic interaction between R318 (arrestin) and E239 (rhodopsin). **e**, The position of the back loop correlates with the interdomain twist angle for simulations of arrestin with its C tail removed, starting from either the inactive (green) or active (purple) state ($R^2 = 0.50$ and $R^2 = 0.58$, respectively; six independent simulations). Back loop position is measured by projecting the coordinates of the back loop onto the vector connecting the crystallographic inactive- and active-state back loop structures (see Methods). **f** Similarly, in simulations of arrestin bound to the receptor core only, movement away from active interdomain twist angles weakly correlated with disruption of the R318–E239 interaction ($R^2 = -0.14$; six independent simulations). Our simulations therefore indicate that interaction between arrestin and receptor at IL3 may control the interdomain twist angle. We speculate that this occurs because the back loop is coupled to the C loop via a set of β-strands. Thus, the receptor is likely to also modulate interdomain twisting by extending the shape of the back loop. When the back loop moves towards its active conformation, its motion appears to couple to the C domain through β-sheet formation with the C loop. Indeed, previous studies have indicated that acidic residues on IL3 might facilitate arrestin engagement. For example, an acidic residue on IL3 of the human luteinizing hormone receptor is critical for binding to arrestin-2 and arrestin-3, albeit to different extents for each[65]. Our simulations support the idea that binding via the IL3 interface could help to trigger arrestin activation. Arrestins 1 to 4 share a conserved basic residue at position 313 (bovine arrestin-1 numbering). A qualitative examination of GPCR sequences reveals that several receptors, including the M2 muscarinic receptor, melatonin receptors, β2AR, A2AR, NTS1R, apelin receptor and H1R, all contain acidic residues at the 5x73–5x75 positions (GPCRdb numbering[66]), which extend into IL3 and may facilitate arrestin activation in the absence of R_P-tail phosphorylation.

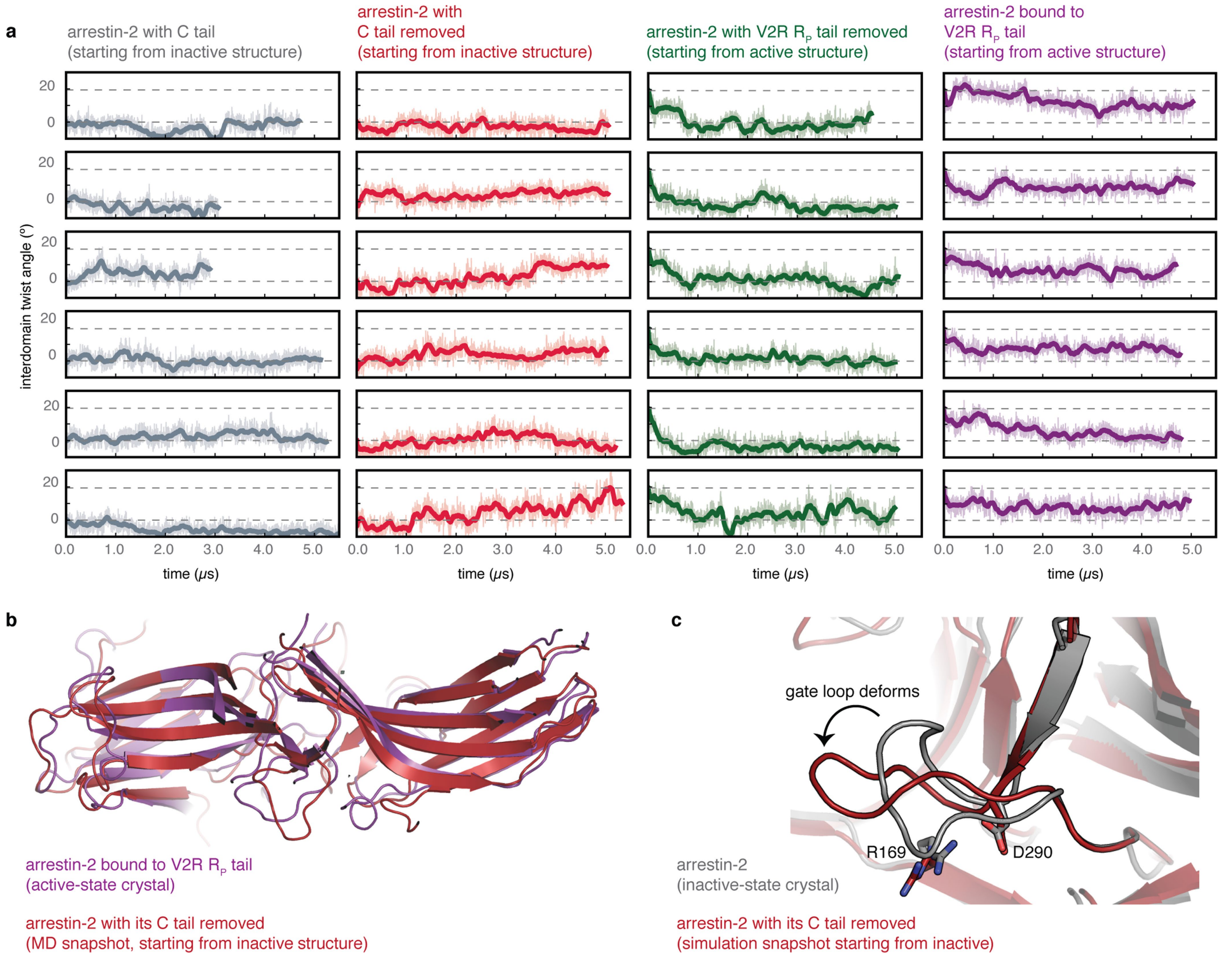


Extended Data Fig. 7 | See next page for caption.

Extended Data Fig. 7 | Gate loop conformation and behaviour of the R175–D296 polar core interaction across arrestin-1 conditions. a, Gate loop conformation (grey) is measured by projecting the coordinates of the gate loop onto the vector connecting the crystallographic inactive and active gate loop structures (see Methods). In simulations where arrestin-1 maintains a stable active interdomain twist, such as in simulations performed in the presence of the R_P tail (right column), the R175–D296 interaction occasionally reforms transiently (blue traces), although the separation of the R175–D296 Cα atoms (red traces) continues to resemble the distance seen in active-state crystal structures. The CHARMM36 force field used here might slightly overstabilize this ionic interaction, increasing the propensity for R175–D296 to reform in these simulations[67]. **b**, Conformations of the polar core and gate loop are tightly coupled. Crystal structures of arrestin-1 bound to its C tail (PDB entry 1CF1) and of arrestin-1 bound to the R_P tail (PDB entry 5W0P) reveal distinct arrangements of the residues in the polar core (D30, R175, D296, D303 and R382) and a surrounding polar network. As described in the main text, the position of D296 is tightly coupled to interdomain twist. In the active state (right), binding of the R_P tail has two effects: first, a phosphorylated serine, S338, engages the gate loop through a direct interaction with K300. In doing so, D296 shifts away from its inactive-state position towards the C domain. In doing so, the interaction between D296 and R175 breaks, disturbing the ionic network of the polar core. Second, S338 also engages R29, which stabilizes the rearrangement of residues in and around the polar core. D30 now engages R175 through an ionic interaction, and D296 is free to interact with other residues, including S308. **c**, Our simulations reveal how the position of D296 is coupled to interdomain twisting. After the gate loop undergoes a conformational change from its inactive conformation to an active conformation in simulations started from the inactive state with arrestin C tail removed, D296 can shift between its inactive and active positions. In snapshots such as the one shown (simulation 8, right), shifting of D296 towards its active position moves a small β-strand (G292–D296), which is connected to a large β-strand (N271–L280) in the C domain. These observations explain the fact that the R175E and D296R mutations—which would force D296 towards its active position by ionic repulsion—cause phosphorylation-independent arrestin activity, whereas the combination of the two mutations, which would maintain the polar core salt bridge between positions 175 and 296, does not[5,68].



Extended Data Fig. 8 | Centrifugal pull-down analysis of fluorescently labelled arrestin mutants. a, Arrestin mutants were mixed with rod outer segment membranes containing rhodopsin (Rho), phosphorylated rhodopsin (RhoP), opsin (Ops) or phosphorylated opsin (OpsP). Rhodopsin samples were illuminated (> 495 nm, 15 s) to obtain activated rhodopsin (Rho*) and phosphorylated activated rhodopsin (Rho*P), and then all samples were centrifuged at 20,800*g* for 10 min. The supernatant was removed, and the pellets were solubilized in loading buffer. Samples were subjected to SDS–PAGE, and gels were stained with Coomassie blue. MW, molecular weight marker kDa. Arrestin migrates slower than rhodopsin or opsin (arrestin (A) and receptor (R) bands are indicated by arrows). As controls, samples of arrestin in buffer alone (NS, nonspecific pull-down in isotonic or low salt buffer) or rhodopsin alone (bkd, background) were centrifuged alongside the other samples. The total amount of arrestin present in each assay (2.25 μg) is indicated in the lanes marked 'Arr'. Arrestin 'cysless' corresponds to the background construct for all fluorescently labelled arrestin mutants (C63A, C128S, C143A, W194F) and is functionally equivalent to native wild-type bovine arrestin-1[61]. Representative gels, cropped to show desired lanes, are shown. Experimental conditions: 1 μM arrestin, 10 μM receptor, 50 μl sample volume; 50 mM HEPES, 130 mM NaCl pH 7 (isotonic buffer) for samples containing rhodopsin, 50 mM HEPES pH 8.5 (low-salt buffer) for samples containing opsin, 20 °C. **b**, Arrestin bands were quantified by densitometry using the program GelQuant.NET v.1.8.2. Each band is expressed as the fraction of total arrestin that was present in each experiment, and bars represent averages from $n = 2$ (arr cysless), $n = 5$ (Rho* and Rho*P, L173F), and $n = 4$ (all other conditions) independent experiments ± s.e.m. Essentially no background density from ROS was present at the molecular-weight range of arrestin. All mutants showed some amount of nonspecific pull-down in the different buffer conditions. Note that this nonspecific pull-down is subtracted from the pull-down data reported in the main text. The fluorescent NBD-labelled arrestin mutants bound to the different receptor variants at similar levels as the cysless arrestin control (Ops < Rho* < OpsP < Rho*P).



Extended Data Fig. 9 | Arrestin-2 undergoes similar fluctuations in simulation as arrestin-1, suggesting a potential common activation mechanism. Simulations initiated from the inactive conformation but with the arrestin C tail removed reached active conformations, and simulations initiated from the active conformation but with the co-crystallized R_P tail removed reached inactive conformations. **a**, Interdomain twist angle as a function of time for simulations of arrestin-2 performed under four conditions: active arrestin-2 bound to the V2 vasopressin receptor C-terminal phosphopeptide (PDB entry 4JQI) with the crystallographic Fab30 fragment removed; active arrestin-2 with the V2R R_P tail removed; and inactive arrestin-2 with its crystallographic C tail present or absent (PDB entry 1G4M). In these simulations, arrestin-2 appears to favour more inactive-like conformations than those seen in the majority of our arrestin-1 simulations, but this may be due to the specific choice of crystal structure from which the simulations were initiated; see Extended Data Fig. 5. Dashed lines represent the inactive and active state interdomain twist angles for the arrestin-2 crystal structures. **b**, Snapshot of an active-like rotational state observed in simulations started from an inactive-state structure with the C tail removed (simulation 63, dark red), overlaid on the active-state structure (purple). **c**, Simulation snapshot from a simulation started from the inactive state with the C tail removed, in which the gate loop moves into an intermediate state (simulation 62, dark red). The absence of a structure of a receptor-bound β-arrestin leaves open the possibility that receptors might bind β-arrestins differently from arrestin-1, and even if the binding mode is similar, the activation mechanism might be different.

Extended Data Table 1 | A non-exhaustive list of experimental studies supporting the hypothesis that arrestin activation depends on receptor engagement of the R_P tail and/or receptor core binding interfaces

Key literature	Observation	Our structural interpretation	Evidence for role of R_P tail, core or both in binding and/or activation
Bouvier et al., *Nature* 1988[69]	Removal of phosphorylable residues on β_2AR C terminus delayed, but did not ablate, receptor desensitization	Arrestin-2/3 still can bind receptor and block G protein coupling, even if R_P tail lacks phosphorylation.	Both (binding)
Ohguro et al., *Prot. Sci.* 1994[70]	Sites that experienced a change in solvent exposure upon rhodopsin binding included residues at the core and tail interfaces and near membrane anchor	Receptor binds to regions at both the tail and core interfaces	Both (binding)
Gurevich and Benovic, *Mol. Pharmacol.* 1997[71]	R175E (arrestin-1) mutant exhibited increased affinity for non-phosphorylated rhodopsin over WT arrestin	Mutations at the R_P tail interface favor conformational changes that increase binding at the core interface (allostery)	Both (binding)
Vishnivetskiy et al., *JBC* 2000[72]	Mutation of hydrophobic residues in the arrestin N-terminus, which binds the arrestin C tail, increased binding to Rho*	Displacing the arrestin C tail (bound at the R_P tail interface) favors conformational changes that increase binding at core interface	Both (binding)
Richardson et al., *J. Neurochem.* 2003[73]	Substance P receptor (SPR) with truncated R_P tail experienced desensitization and internalization to similar extents as wild type SPR and retained some affinity for arrestin-3	Receptor without R_P tail can still bind arrestin and induce arrestin-mediated processes	Core (activation)
Jala et al., *JBC* 2005[23]	When phosphorylable residues of leukotriene B4 receptor were eliminated, the receptor still engaged arrestin 2 or 3, enabling arrestin localization to membranes and receptor internalization. These phosphorylation-deficient receptors failed to undergo endosomal localization	While the R_P tail appears to favor arrestin interactions with the receptor that promote endocytosis, absence of R_P tail does not prevent receptor-dependent arrestin localization to membrane. R_P tail binding may increase lifetime of arrestin-receptor complex, enabling endocytosis.	Both (activation)
Hanson et al., *JBC* 2006[74]	Mutations at the interdomain interface increased binding to both Rho* and inactive RhoP; mutating basic residues in the N domain impaired binding to RhoP more than it impaired binding to Rho*P	Breaking contacts at the interdomain crevice, near the core binding interface, increased binding at the R_P tail interface, suggesting that changes at the core interface favor changes at the tail interface	Both (binding)
Shukla et al., *PNAS* 2008[75]	AT1aR with truncated R_P tail induced conformational changes in arrestin-3, as measured with BRET, similar to those induced by full-length AT1aR.	BRET assays likely report on displacement of arrestin's own C tail. Thus, receptor core alone appears capable of favoring C tail displacement	Core (activation)
Gimenez et al., *JBC* 2012[76]	Phosphorylation-deficient M2R bound arrestin-2/3, while phosphorylation-deficient β_2AR experienced reduced binding, as measured with BRET	Core binding appears to play a role in M2R-mediated arrestin activation, while β_2AR appears to depend more upon R_P tail–mediated activation	Both–dependent on receptor type (binding)
Vishnivetskiy et al., *JBC* 2013[77]	Mutation of residues in the arrestin-1 middle loop (Q133–S142) increased arrestin-1 binding to OpsP as well as to Rho*	Breaking interdomain contacts (through mutation) near the core interface increases binding at the R_P tail interface	R_P tail (binding)
Yang et al., *Nat. Comm.* 2015[24]	Binding of V_2R- and β_2AR-derived phosphopeptides induced conformational changes in arrestin detected by ^{19}F-NMR	R_P tails with distinct phosphorylation patterns can induce different arrestin conformations and favor distinct downstream effects associated with arrestin activation	R_P tail (binding and activation)
Kumari et al., *Nat. Comms.* 2016[10]	An R_P tail–engaged form of β-arrestin 1 experienced endocytosis and mediated ERK signaling	Binding to a highly phosphorylated R_P tail stimulates arrestin activation	R_P tail (activation)
Thomsen et al., *Cell* 2016[12]	EM images of a GPCR supercomplex revealed that arrestin and G protein can simultaneously bind a single GPCR *in vitro;* also, arrestin-3 and Gαs colocalized with GPCRs in endosomes	Since the G protein engages the receptor core, arrestin likely simultaneously engages the GPCR by binding at the R_P tail interface	R_P tail (binding)
Lee et al., *Nature* 2016[25]	Swapping R_P tails of certain receptors altered intracellular arrestin-3 BRET signals and resulted in different cellular trafficking	Binding of R_P tails with different extents of phosphorylation may favor distinct cellular outcomes by altering the lifetime and conformational signature of arrestin	R_P tail (activation and binding)
Cahill et al., *PNAS* 2017[13]	Mutation of the finger loop, a core-binding element in arrestin, in a chimeric β_2V_2R did not prevent receptor internalization or signaling	Ablating binding via the core interaction does not prevent arrestin-related processes (internalization, eg.), implying a role for R_P tail in promoting arrestin-mediated signaling	R_P tail (activation and binding)
Kumari et al., *Mol. Biol. Cell.* 2017[11]	A V_2R mutant lacking ICL3 could still engage clathrin and other components of the ERK signaling module	Ablating another portion of the core interface did not prevent arrestin activation, implying a role for R_P tail	R_P tail (activation and binding)
Jung et al., *PNAS* 2017[22]	Phosphorylation independent, transient binding between M1R and arrestin-3; arrestin-3 conformation detected by FRET	Weaker of two binding modes is phosphorylation independent; the more stable mode may involve IL3 phosphorylation at S228, leading to further engagement at the core interface	Core (activation and binding)
Eichel et al., *Nature* 2018[25]	Agonist-stimulated receptors with C-terminal truncations still experienced arrestin-mediated membrane localization and receptor internalization	Receptor core alone can trigger certain changes in arrestin, but R_P tail required for receptor endocytosis	Both (activation)

Published data that support the hypothesis that the receptor core and R_P tail independently promote arrestin activation[10–13,22–25,69–77]. Note that the studies supporting a role for the core in arrestin activation do not preclude a role for the R_P tail, and those supporting a role for the R_P tail in arrestin activation do not preclude a role for the core. Furthermore, not all studies directly distinguish between interactions that favour binding (that is, increase receptor–arrestin affinity) and those that favour arrestin activation (that is, conformational change in arrestin).

BLOG Personal stories and careers counselling **http://blogs.nature.com/naturejobs**

TWITTER Discover career resources and tips **go.nature.com/2xyqla3**

NATUREJOBS For the latest career listings and advice **www.naturejobs.com**

ADAPTED FROM PLANET FLEM/GETTY

LAB LIFE

Learn to lead

Laboratory heads often aren't properly trained for their job. Here's what they need.

New principal investigators (PIs) tend to be thrust into a role that includes administration and management, budgeting, lab infrastructure, mentoring and — perhaps most difficult — shepherding junior scientists to achieve their full potential. Many researchers receive little training in leadership skills to help them adapt. And the problem is not limited to just new PIs. In *Nature*'s 2017 survey of PhD students, one-quarter of the respondents were dissatisfied with their adviser's guidance on research, and only one-third said that their supervisors gave useful advice about non-academic careers.

Nature asked leadership experts what lessons other industries can offer to help scientists become more-effective PIs.

NATURE.COM
For more on science in culture see:
nature.com/booksandarts

PETER HIRST

Treat science like a business

Associate dean of executive education at the Massachusetts Institute of Technology (MIT) Sloan School of Management in Cambridge.

Our programmes at Sloan on management and leadership attract people from fields such as financial services, government, engineering and the life sciences. In my view, it would help scientists to think about the parallels between science and the corporate world: understanding what the product is, who the customers are and whether you're creating value for them. In scientific research, your 'customers' aren't necessarily paying you. But a lot of scientists would say, "My customer is the research community; I'm helping to advance this field."

There are ways to measure the impact of your work and to improve the return on resources you're putting in. For example, you could measure the number of citations, publications or revisions needed to get a paper accepted. Once you have your scorecard, you can look for opportunities to make sure you're focusing resources — your own attention, your students' efforts and lab equipment — to improve those measurements.

Techniques developed for the manufacturing industry can help you to visualize where things have become stuck. Think of a lab ▶

as manufacturing science. Presumably there's some output you're trying to produce. You can apply these approaches to get more throughput.

To use these techniques in a lab, identify every activity or piece of work and write it on a Post-It note. On a large wall or whiteboard, set out the process that tasks flow through. All participants can refer to the visualization on an ongoing basis; it's a continuous representation of the work happening. Some tasks will depend on the outputs of others; task B can't be done until task A has been completed. Maybe team A is characterizing an organism and handing that knowledge on to team B.

> **"It's valuable for scientists to learn about business concepts."**

You could use traffic-light colour-coding to highlight something that is stuck or has a high error rate: green means everything's going as it should, yellow means we need to keep an eye on it and red means it needs work. Say you want to modify a piece of equipment to improve its performance. But you discover that you need a part that has to be manufactured specially, so the task becomes a yellow flag. Then you find out that it will take six months to make the part, so the project becomes a red. At that point, you seek insight from throughout the organization; for instance, a colleague in an adjacent lab might have had a similar problem and be able to suggest an alternative solution.

It's valuable for scientists to learn about business concepts. The unifying entity is human beings; whether you're leading a business or an academic lab, you depend on getting people to work together to achieve things.

Scientists could ask faculty members in their university's business or management school whether the institution has courses on these techniques (see 'Training options'). Or they could request peer coaching and say, "Your work looks interesting, and I'm curious if I could use these ideas in my lab. Would you help me to understand it?" Investing in your skills will pay dividends for you and those around you.

ADAPTED FROM PLANET FLEM/GETTY

SEN SENDJAYA

Lead by serving others

Leadership researcher at Swinburne University of Technology in Melbourne, Australia.

I have trained people on an approach called servant leadership, which involves serving those whom you lead. Many leadership styles focus on manipulating followers to go the extra mile so that the employer can squeeze more effort out of them, and the end game is to increase the bottom line. Servant leadership is different because it focuses on followers' holistic development and puts their needs above those of leaders.

For example, instead of saying "I expect you to have the report written by Monday at 9 a.m.," a servant leader would say, "What do you think of this expectation? Would you be able to fulfil it?" They might ask whether there is anything going on professionally or personally that would hinder productivity and that the servant leader can help with.

Another distinctive thing about servant leadership is a 'hunting licence': the leader tells followers that they can ask difficult questions and disagree. Servant leaders are willing to be vulnerable and admit mistakes. The key is to communicate this quite regularly because it's human nature to not openly disagree with people above us. Leaders need to remind followers, "I'm not infallible, so you need to tell me if I'm doing something out of line."

Servant leadership is applicable to scientists because people who work in labs are highly creative. Creative people tend to be more autonomous and self-motivated than the average employee; they often don't want to be told what to do. So PIs might not be able to rely on traditional influencing techniques such as hierarchy-based authority.

My team and other researchers around the world have conducted studies on servant leadership in sectors such as manufacturing, hotels, restaurants, airlines, financial services and government. Servant leadership has been shown to boost employee commitment, team-members' performance, intention to stay with the organization and the performance of the company, compared with other leadership approaches. The approach is starting to become more mainstream.

Business practitioners often have difficulties with servant leadership because they think they'll be treated like a doormat. But servant leadership does not operate out of weakness or lack of self-respect. Only those with a secure sense of self can serve others through their leadership.

SUE HEWITT

Make time for training

Proprietor of Develomenta, a training consultancy in Denbigh, UK.

We shouldn't put scientists in leadership positions without giving them support. Employees of many corporate organizations and the UK civil service (national and local government workers) often receive leadership training through workshops and courses, for instance. In academia, however, there is less of an expectation that someone who progresses to a management role will be supported with leadership development.

As a PI, you might avoid seeking leadership training because you don't feel you need it, or because you're too busy. But you don't have time not to do it. For example, if you learn about delegation, think about how much more effective you will be when everyone in the lab is helping you to do things, instead of doing them by yourself.

I teach scientists how to coach people, which means having a conversation to help staff or students to solve their own problems. The main coaching model is called GROW, which stands for goal, reality, options or obstacles and will or way forward. It's a simple technique for asking questions to identify the issue, the current reality, the options and what the person is going to do. For many PIs, it takes a great weight off their shoulders when they realize that they don't have to solve everyone's problems.

For example, if a student thinks that their results aren't good enough for a conference presentation, you could explore what 'not good enough' means. You might find that the student is unduly anxious about the quality of the work when, in fact, the more important thing is for them to get feedback on their initial findings from people at the conference. Other meeting attendees who have used a similar experimental technique could then talk to the student about how they saw the same problems or improved their data. As the PI, you might not know exactly how to improve the results if the student is using a new technique. But by presenting the work and getting their own feedback, a student can take more ownership of the project.

Many PIs feel that they aren't able to give students advice on non-academic careers, such as scientific publishing. Most PIs have never done that type of work. But they should be able to say: "I've submitted a paper to this journal; maybe I could ring them and ask someone to talk to you." There are many ways to help. It's difficult to provide training to PIs on this issue because of the huge variety of careers that people can choose, but PIs can bring students together with the university's careers service or with an organization such as hfp consulting in Heidelberg, Germany, which offers professional development workshops for scientists.

TRAINING OPTIONS

Available courses

Here are some organizations that offer leadership programmes geared towards scientists.

- hfp consulting in Heidelberg, Germany, runs a variety of professional-development workshops for scientists at all levels.
- Cold Spring Harbor Laboratory in New York runs annual interactive three-day workshops on leadership in bioscience.
- The University of California, San Francisco, offers a 16-hour course on scientific leadership and management skills that targets people hoping to lead research groups
- The Technical University Berlin, the Humboldt University of Berlin and the Free University Berlin have joined forces to offer their female researchers a programme named ProFiL, which combines mentoring, seminars and training sessions.

It's also important for female PIs to have access to women-only leadership training. I run courses for female scientists, and cover a lot of the same material as standard leadership courses, but with a gender slant. For example, the courses ask about their experience in giving or receiving feedback as a woman, and help them to be more confident. Because we provide an environment in which all the participants and many of the trainers are female, women's voices are easier to hear. Dealing with feelings about coming back to work after maternity leave, for example, is openly talked about. Female PIs should ask their universities or funding bodies whether they provide leadership training for women and make a case for them to do so. We already have the evidence that there aren't enough senior women in science.

KEN INGRAM

Embrace the uncertainty

Head of practice at Roffey Park Institute, a leadership-development organization in Horsham, UK.

We find that experts often have trouble when they get to leadership positions. By 'experts,' I mean anyone who has been successful because of their skills or knowledge, whether they're, say, a nurse, an accountant or a scientist. They often experience 'impostor syndrome', and think, "I don't feel competent because this is a new skill set, and it's not what I relied on to get me here." It can be quite an anxiety-producing time.

Many people who are used to being in a position of expertise feel uncomfortable with the ambiguous nature of managing people. When they were in specialist roles, there were certain procedures and outputs expected. For a web designer, success is clear: a great webpage that attracts people. But when that person becomes a leader, they're no longer creating the webpages; they're helping other people to do so, and that can be difficult. They often instinctively revert to what they feel most comfortable with — for example, doing the web design themselves. But it's important to let team members come up with their own ideas and have some autonomy, because that's how people learn.

Experts often seek a single truth that is pure fact, but leadership is greyer than that. For instance, one team member might feel that they are always interrupted in meetings. They might feel bullied, and see the person talking over them as arrogant, whereas the person behaving that way might say, "I was passionate — I'm an extrovert, so I speak what I'm thinking." A leader with a 'single-truth' mindset might dismiss one person's point of view. But that's not helpful in an interpersonal situation, because everyone experiences the world differently. Instead, the leader should help both parties to see the other person's point of view.

KATE MACMASTER

Cultivate self-awareness

Programmes director at the Peter Cullen Water and Environment Trust in Canberra, Australia.

Our Science to Policy Leadership Program is aimed at emerging and current leaders in Australia who work in water and environmental management. A lot of them are scientists. Part of the process we take them through is increasing their awareness of personal purpose and values. Participants are asked to take a psychometric test called DiSC, which stands for dominance, influence, steadiness and conscientiousness. They answer multiple-choice, scenario-based questions, and the test assigns them a broad personality type. For example, dominant people tend to focus on getting results quickly and succeeding, and conscientious people are more concerned with understanding details and working carefully.

The important thing to remember is that you are not necessarily just one profile; you can have parts of all four. But the test does give your general preference in any situation, and the aim is to notice and be aware of it. For example, I might be required to lead my group and make high-level decisions. If I know my personality preference is to get stuck in the details and micromanage, and I notice myself acting that way, I can reflect on the spot and behave differently: "I'm going to stop doing that, go back to high-level strategy and trust my team."

My advice to principal investigators is to spend some time with your team identifying your shared purpose and values. You might decide your values are hard work, respect, honesty and integrity. Then, if a project gets off-track because people aren't, for instance, working hard or acting with integrity, you can bring the group back together and say, "Let's revisit our values and talk about how we're doing." Once the team is aligned with a clear purpose, you find increased trust in the group. Trust is the foundation for any high-performing team, whether they're scientists, politicians, schoolteachers or retail assistants. ■

INTERVIEWS BY ROBERTA KWOK

Interviews have been edited for clarity and length.

WHEN NAIN CAME TO SHIRIN'S DOOR

Stories from the stars.

BY FILIP WILTGREN

The first time Nain came to Shirin's door, he was a young man, full of promise. They dallied over green mint tea and the date-crumb cakes Shirin's mother had baked. The sun filtered through the grape leaves over the veranda.

"Come with me to the stars," Nain said. "With you at my side, there is nothing I can't do."

"What about my family?" Shirin said. "What about my dancing and my friends?"

Nain let a server refill his mug of tea. The leaves reflected from the server's chromed carapace, making the robot look like part of nature. Hand-beaten carapace, hand-driven clay mugs. Shirin's family was respected, their daughter young and comely.

"It will only be two years," Nain said.

"For you."

She didn't say 'for us'.

The second time Nain came to Shirin's door, he was a young man, full of stories. He ate Shirin's hand-baked date cakes, and told of paragliding in the volcanoes of Radhir, of floating in space among the moons of Nuova Sol, a billion stars shining before his faceplate.

Shirin listened politely, her rheumy hands folded in her lap, her children and grandchildren around her. In the short breaks when Nain stuffed his face with cake, she told of dancing in the festivals, of her choreography and her children. Nain listened, but his heart wasn't in it, and her children wanted more stories of faraway planets, so in the end she fell silent.

The third time Nain came to Shirin's door, he asked to speak with Riva. Riva was a lithe beauty, with hair the colour of midnight and eyes that sparkled like a billion stars.

Shirin let Nain into her house, and watched, worry lining her face. The two youths talked on the veranda, below the cover of the grape vines. Riva laughed, and flirted, and each laugh was a spear in Shirin's heart.

The children's talks grew animated, then heated, and in the end Riva stormed away, while Nain bowed his way out.

"Did he ask you to go away with him?" Shirin asked. Her granddaughter nodded, a motion proud in its stiffness.

"Why did you decline?" asked Shirin.

"That boy's heart is a billion miles away," Riva said, and Shirin nodded.

NATURE.COM
Follow Futures:
@NatureFutures
go.nature.com/mtoodm

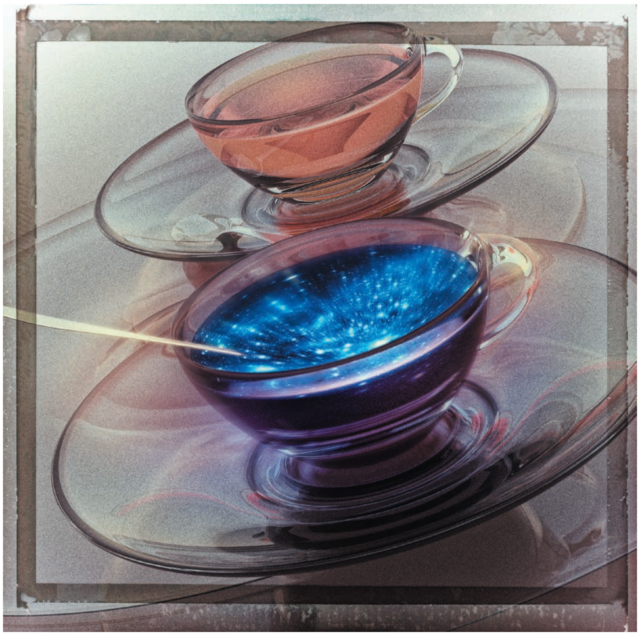


ILLUSTRATION BY JACEY

The fourth time Nain came to Shirin's door, he carried a force-shield on his hip, and a pair of nanoscale gun-bots riding on his shoulders. He found the house abandoned, holes blasted in its walls, grape vines burned stumps.

"Hey," he said, grabbing a nearby squatter. "The family who lived there, where did they go?"

The squatter shrugged, an elaborate gesture involving the rolling of his eyes and the wiggling of his fingers.

"Some went there," he said, pointing to the distant mountains. "Some went there —" the squatter pointed at the sky, "— and some went there." He pointed to a row of small mounds along one of the broken walls.

Nain went over to the graves, but all the stones were unmarked.

The fifth time Nain came to Shirin's door, he strode down a wide avenue, face lined, gold piping on his shoulders, gold braids across the midnight-blue of his uniform. He banged on the reinforced steel door, which slid aside to reveal a human servitor, a middle-aged woman, dressed wholly in white, with a slave-collar around her neck.

"There was a family who lived here," he said. "Where are they now?"

"The master's family has always lived here, even before the wars," the servitor answered, bowing, and Nain invited himself in.

He looked at the frescoes on the walls, tall men in martial poses, tanks and walkers at their back.

"I knew your greatmother," he told the master of the house, a handsome man with an easy smile and friendly laugh. The man waved the information away.

"Ancient history," he said. "Come, tell us of the stars."

Nain bowed his head, and obliged, then went back whence he'd come.

The sixth time Nain came to Shirin's door, his steps were heavy and his hair was grey.

The armour on the door was gone, the house had gained an extra floor, and the veranda had been enclosed in glass. Apple trees grew in the courtyard, and a gaggle of young girls chased each other beneath them.

"Have you lost your way, grandfather?" one of them asked.

Nain sank down to her level, leaning an elbow on his knee and supporting himself with a hand on the grass.

"What's your name?" he asked.

"Sahrin," the girl answered, and with the self-assurance of the truly young added, "Ma says its after one of our family's firstparents."

"Sahrin," Nain repeated, a crooked smile bisecting his face. "I think I may have known your greatmother, once upon a time."

"Will you tell me about her?"

Nain drew breath, but his answer was cut short. A man and a woman had come out on the porch.

"Don't bother the gentleman," chided the woman.

"No bother," said Nain. "I was just about to tell a story."

"Well then," said the man, "a story freely shared is a treasure to the heart. Would you come in and share it with all of us?"

And Nain entered, and told his story, and when it was time to leave, he stayed. ■

Filip Wiltgren *is a communications officer at Linköping University by day and a writer, game designer and worrier by night. He lives in Sweden with his wife and children.*